RECPAD'20

26th Portuguese Conference on Pattern Recognition

October 30, 2020
Remote Event

Proceedings

UNIVERSIDADE DE ÉVORA

APRP

Proceedings of RECPAD 2020
Escola de Ciências e Tecnologia
Universidade de Évora
2020

26th Portuguese Conference on Pattern Recognition

https://recpad2020.uevora.pt

Contents

Poster Session 2

Poster Session 3

107 Author Index

Preface

This volume contains the collection of papers accepted for RecPad 2020. RecPad is the annual Portuguese Conference on Pattern Recognition, promoted by APRP (Portuguese Association for Pattern Recognition). It is a one-day conference that aims to promote the collaboration between the Portuguese scientific community in the fields of Pattern Recognition, Image Analysis and Processing, Soft Computing and related areas. Topics include (but not limited to):

- Statistical, structural, syntactic pattern recognition;
- Neural networks, machine learning, data mining;
- Discrete geometry, algebraic, graph-based techniques for pattern recognition;
- Signal analysis, image coding and processing, shape and texture analysis;
- Computer vision, robotics, remote sensing;
- Document processing, text and graphics recognition, digital libraries;
- Speech recognition, music analysis, multimedia systems;
- Natural language analysis, information retrieval;
- Biometrics, biomedical pattern analysis and information systems;
- Special hardware architectures, software packages for pattern recognition.

On its 26th edition, RecPad 2020 was organized by the University of Évora and held as a remote event on October 30, 2020. In this edition **53 papers were accepted** for poster presentation and the best 5 submissions were selected for oral presentation. Besides the oral and poster sessions, RecPad 2020 also featured:

- an invited talk by Dr. Hubert Shum, titled "Machine Learning for Human Data Modelling and Analysis";
- prizes for the best oral presentation and best poster sponsored by DECSIS.

We would like to express our appreciation to all the authors and members of the scientific and organizing committees which were a key contribution to the success of this conference - our big Thank You!

Sponsors & Partners

Committees

Organizing Commitee

Teresa Gonçalves, Universidade de Évora
Luís Rato, Universidade de Évora
Pedro Salgueiro, Universidade de Évora
Miguel Barão, Universidade de Évora
Eduardo Medeiros, Universidade de Évora

Scientific Committee

Ana Filipa Sequeira, INESC TEC
Ana Mendonça, Universidade do Porto (FEUP), INESC TEC
Ana Rebelo, Universidade Portucalense, INESC TEC
António Cunha, Universidade de Trás-os-Montes e Alto Douro, INESC TEC
Armando Pinho, Universidade de Aveiro, IEETA
Augusto Silva, Universidade de Aveiro, IEETA
Beatriz Sousa-Santos, Universidade de Aveiro, IEETA
Bernardete Ribeiro, Universidade de Coimbra, CISUC
Carlos Ferreira, INESC TEC
Catarina Silva, Universidade de Coimbra, CISUC
César Teixeira, Universidade de Coimbra, CISUC
Diogo Pratas, Universidade de Aveiro, IEETA
Fernando Monteiro, Instituto Politécnico de Bragança
Francesco Renna, Universidade do Porto (FCUP), Instituto de Telecomunicações
Hélder P. Oliveira, Universidade do Porto (FCUP), INESC TEC
Hugo Silva, Instituto de Telecomunicações
Inês Domingues, Instituto Superior de Engenharia de Coimbra, Centro Investigação IPO
Jaime Cardoso, Universidade do Porto (FEUP), INESC TEC
Joana Costa, Instituto Politécnico de Leiria, CISUC
João Carlos Neves, Instituto de Telecomunicações
João Rodrigues, Universidade do Algarve
João Sanches, Universidade de Lisboa (IST), ISR
Joel P. Arrais, Universidade de Coimbra
Jorge Marques, Universidade de Lisboa (IST), ISR

Jorge Oliveira, Instituto de Telecomunicações
Jorge Santos, ISEP
Jorge Torres, Instituto de Telecomunicações
José Saias, Universidade de Évora
José Silva, Academia Militar
Luís Alexandre, Universidade da Beira Interior
Luís Rato, Universidade de Évora
Luís Teixeira, Fraunhofer Portugal AICOS
Mário Antunes, Instituto Politécnico de Leiria (ESTG), INESC TEC
Mário Figueiredo, Universidade de Lisboa
Miguel Barão, Universidade de Évora
Miguel Coimbra, Universidade do Porto (FC)
Nuno Rodrigues, Instituto Politécnico de Leiria (ESTG), Instituto de Telecomunicações
Paulo Salgado, Universidade de Trás-os-Montes e Alto Douro
Pedro Pina, Universidade de Lisboa
Pedro Salgueiro, Universidade de Évora
Petia Georgieva, Universidade de Aveiro
Tânia Pereira, INESC TEC
Teresa Gonçalves, Universidade de Évora
Thomas Gasche, Academia Militar
Verónica Vasconcelos, Instituto Superior de Engenharia de Coimbra (IPC)

Invited Speaker

Hubert P. H. Shum - Durham University
http://hubertshum.com

Biography

Dr Hubert P. H. Shum is an Associate Professor in Computer Science at Durham University. Before this, he worked as the Director of Research/Associate Professor/Senior Lecturer at Northumbria University, a Postdoctoral Researcher at RIKEN Japan, and a Research Assistant at the City University of Hong Kong. He received his PhD degree from the University of Edinburgh, his Master and Bachelor degrees from the City University of Hong Kong. He led funded research projects as the Principal Investigator awarded by EPSRC, the Ministry of Defence (DASA) and the Royal Society. This facilitated him to develop his research team and to collaborate with international researchers from the UK, China, France, Japan and India. To engage the academic and industry networks, he led his team hosting important conferences such as BMVC and ACM SIGGRAPH Conference on MIG. Contributing to the research community, he has served as an Associate Editor for Computer Graphics Forum, a Guest Editor for International Journal of Computer Vision, and a Program Committee member in 15 conferences such as CVPR, Eurographics, Pacific Graphics. He has published over 100 research papers in the fields of computer graphics, computer vision, motion analysis and machine learning, particularly focusing on the modelling of human-related data. More information can be found on http://hubertshum.com.

Title:

Machine Learning for Human Data Modelling and Analysis

Abstract:

Taking advantage of the recent advancement in machine learning and the availability of big data, computers have become smarter than ever in understanding complicated data. In this talk, I will focus on the modelling and analysis of human data, which can be represented in formats such as images, video, 3D movement and surfaces. Such data is core to a wide spectrum of research fields including computer vision (e.g. action recognition, pose estimation, 3D reconstruction), computer graphics (e.g. character animation, crowd simulations) and biomedical engineering (e.g. diseases diagnosis, motion analysis). Modelling human data effectively is a challenging problem as it is high dimensional in nature and diverse in representations. I will talk about how machine learning techniques can be used to take on the challenge to come up with novel models that enable robust applications. In particular, I will discuss how state-of-the-art deep learning provides a powerful framework for large-scale human data modelling and analysis. Finally, I will share some insights into future research opportunities and interesting research directions in this area.

Training Convolutional Neural Networks to be Background Invariant

Ricardo Cruz
ricardo.p.cruz@inesctec.pt
Jaime S. Cardoso
jaime.cardoso@inesctec.pt

INESC TEC
Faculty of Engineering, University of Porto
Portugal

Abstract

Convolutional neural networks have been shown to be vulnerable to changes in the background. For example, a CNN trained using objects on top of a blue background often performs terribly when evaluated using a green background. The proposed method is an end-to-end method that augments the training set by introducing new backgrounds during the training process. The novelty is that these backgrounds are created on-the-fly using a generative network that is trained as an adversary to the model. The adversary dynamics ensures that the model has seen a wide range of backgrounds. The method is experimented with using MNIST and Fashion-MNIST as test cases.

1 Introduction

Possibly due to the fact that neural networks learn from static images, and so do not have to deal with depth as us humans, they have a brittle understanding of what an object is and are vulnerable to changes in the background – for example, when there is a mismatch in the background between the training and test sets, performance degrades terribly, as exemplified by Figure 2. In that case, the classifier is trained with digits in a clean, white background (a trivial task) and then evaluated with digits inserted in diverse backgrounds.

These disparities in background between training and testing set have not been studied in detail. There is one work that uses an attention mechanism but only avoids some artifacts, such as irregular borders [2].

Generative adversarial networks (GAN) generate realistic images through a min-max problem whereby two models (generator vs discriminator) try to optimize a given loss function in the opposite direction [1]. This work is loosely inspired by this dynamic. A generator is proposed to augment the training set by producing backgrounds that purposefully have an adverse effect on the performance of the target model, making the target model more robust as a result. To introduce the new background, the object must first be segmented therefore a third neural network that is trained in an unsupervised manner.

2 Related Work

Literature exists in predicting classifier confidence for dataset shifts so that changes in the background could be detected. However, making the classifier itself robust to changes in the background seems to have been the subject of little study. One work proposes an attention mechanism to avoid artifacts, particularly irregular borders, from influencing the classifier [2]. Two classifiers are used: a *global* CNN, G, and a *local* CNN, L. The proposed method works by having G find the bounding box of the relevant object in order to create a cropped version of the image, and then use L to classify the cropped version.

That attention mechanism works in two phases. Firstly, G is trained to classify the entire image x. Then, a truncated version G^T is used to obtain activation maps and find a bounding box around the object so that a function f produces a cropped image x'. Finally, L is then trained using

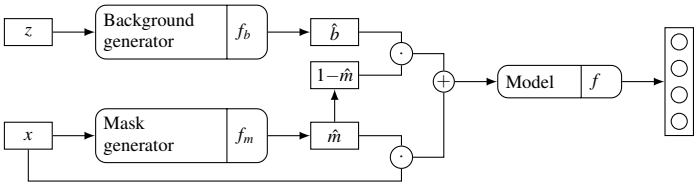
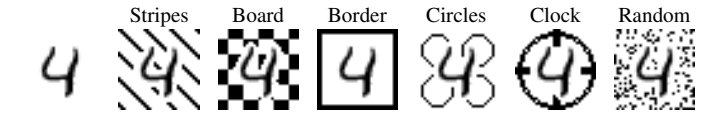


Figure 1: Proposed adversarial background augmentation during training.



	Stripes	Board	Border	Circles	Clock	Random
Traditional	38.0	24.3	61.4	32.9	19.7	11.2
Proposal	92.3	76.8	93.1	93.7	70.8	86.2

The model is a CNN with VGG blocks as detailed in Section 4, trained for MNIST. Accuracy values for the entire testing set when different backgrounds are used.

Figure 2: Background change can produce wild disparate accuracies (%).

the smaller image x' [2]. To then predict the class y of the image x, this chained process can then output the class $\hat{y} = L\big(f(G^T(x),x)\big)$.

Two disadvantages are immediate: (i) L operates on a rectangular cropped version of the image and therefore is still influenced by artifacts that remain inside that rectangle, and (ii) model G is still influenced by artifacts because it did not have the benefit of being trained against the artifacts. While such artifacts are not presented in the training set, they could be generated in a controlled fashion, as the method next proposed.

3 Method

The goal is to (during training) be able to place the object in a multitude of contexts (backgrounds), facilitating the learning of robust representations, focused on "what" the object is rather than "where" the object is. We propose to adopt adversarially generated backgrounds to promote the learning of strong representations. However, the insertion of adversarial backgrounds in the image cannot be allowed to destroy the concept (class) one is trying to learn. Since the spatial delineation of the object is unknown, we propose to learn, simultaneously with the recognition, the segmentation mask. This mask is used to inject the adversarial background only in the non-object pixels.

Model: A model f is optimized to minimize a loss $\mathcal{L}(y, f(x))$ using an image x as input with label y as the ground-truth. This image is subject to data augmentation through the process illustrated in Figure 1.

The framework is agnostic of the task and other losses could be used for different tasks: regression, semantic segmentation, reinforcement learning, etc. In these experiments, classification was performed using cross-entropy,

$$\mathcal{L}(y,\hat{y}) = \sum_{i=1}^{N} y_i \log \hat{y}. \tag{1}$$

Mask generator: Firstly, a model f_m is trained to produce a mask $\hat{m}$ using a sigmoid activation function to ensure $\hat{m} \in [0,1]$ so that it can be used to segment the image through a element-wise product, $x' = x \odot \hat{m}$. The model f_m can be optimized in an unsupervised fashion by finding the best mask that minimizes the previous loss, $\mathcal{L}\big(f(x \odot f_m(x)), y\big)$. To help prevent the mask from including background, a term $\mathcal{L}_A$ is used to constraint its size

$$\mathcal{L}_A(\hat{m}) = \max(0, A(\hat{m}) - a), \tag{2}$$

where A approximates the percentage of the area of the mask by computing $A(\hat{m}) = \frac{1}{wh}\sum_{x,y}\hat{m}_{x,y}$, and a is the average area for the object given as an hyperparameter ($a = 0.2$ is used in all experiments). For better performance, after model f_m has been trained, a non-differentiable transformation T can henceforth be applied to further improve the segmentation. For example, a threshold t, $T(\hat{m}) = \mathbb{1}_{x,y}(\hat{m}_{x,y} \geq t)$, can be chosen using Otsu's method or the a-quantile such that $A(\hat{m} \geq t) = a$.

Notice that the mask generator being background invariant is unimportant since it is only used during training and on training images. A typical architecture for the mask generator would be a U-Net [5]. For better results, the mask could be provided by the user through manual segmentation.

Background generator: Secondly, the background is generated by a neural network f_g that transforms noise z into a background $\hat{b}$ image. Unlike the others, this model is trained to *maximize* the loss $\mathcal{L}$. The generator focuses on producing backgrounds or artifacts that could potentially adversely affect the output of the model.

In the case of MNIST and Fashion-MNIST which are monochrome, the background generator could "cheat" by producing a background with the same color as the object, thus obfuscating the object. In these cases, a constrain was added in the form of the additional regularization $\mathcal{L}_{B_A}(\hat{b})$ term which is added to disallow the background from filling over half the pixels, $\mathcal{L}_{B_A}(\hat{b}) = \max\left(\frac{1}{N}\sum_{i=1}^{N}\hat{b}_i - 0.5, 0\right)$.

Overall dynamic: All in all, the min-max optimization problem can be summarized as

$$\min_{f,f_m}\max_{f_b}\sum_{i=1}^{N}\mathcal{L}\left(f\left(\hat{m}_i \odot x_i + \hat{b}_j \odot (1-\hat{m}_i)\right), y_i\right) + \mathcal{L}_A(\hat{m}_i) + \mathcal{L}_{B_A}(\hat{b}). \quad (3)$$

Notice that, while the optimization problem was inspired by GANs, this is not a GAN framework, there is no discriminator used. Also, while this problem could be optimized end-to-end, we have performed this optimization in three stages: (i) train model f, (ii) train mask generator f_m, (iii) train both model f and its adversarial background generator f_b. Training in stages is useful for debugging and fine-tuning, but also it allows applying non-differentiable transformations on top of f_m such as thresholds to help produce more realistic masks.

4 Experiments

MNIST [3] and Fashion-MNIST [6] are artificially enhanced by introducing backgrounds as illustrated in Figure 3. This enhanced versions are used only for *testing* purposes, while the original unmodified dataset is used for *training*. The idea is to see how well the model performs when background textures are introduced.

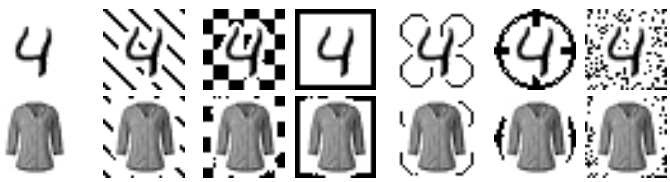
Table 1 summarizes the results showing the proposed method (proposal) to have become background invariant. Interestingly the attention mechanism results only negligibly improve on the baseline classifier. This mechanism works by cropping the image and, not surprisingly, it was found to perform best in the border case (with over 50% accuracy); still, this result was worse than the proposal.

To better understand the impact of changes in the background, let us vary the rate of the random parameter from the previous Figure 3 (g). In Table 2, a Bernoulli distribution is used for the background with varying parameter values, as illustrated in the images. While the baseline naturally produces better results for the unchanged image, as the rate is increased, the drop in baseline's performance is fathomed while the proposal drops more smoothly.

Furthermore, to better understand what could be improved on the framework, the mask generator is changed so that a manual segmentation is used instead of a neural network, and also the background generator is changed to produce noise instead of trained adversarially (see Table 3). Two conclusions are apparent: (a) the fact we train the mask generator in an unsupervised fashion means that the masks are imperfect which greatly influence performance, (b) using noise as the background is not sufficient to avoid the network being fooled by more intricate patterns as those in the testing set (Figure 3).

5 Conclusion

This work was fomented by previous work where the goal was to train a drone using a dataset that was easier to acquire indoors (inside a studio) rather than outdoors where it was going to be used, because it dealt with electricity insulators [4].



(a) Original (b) Stripes (c) Board (d) Border (e) Circles (f) Clock (g) Random

Figure 3: Backgrounds introduced for MNIST and Fashion-MNIST.

Table 1: General Results (Validation Accuracy in %)

Method	Stripes	Board	Border	Circles	Clock	Random	Avg
	MNIST						
Traditional	38.0	24.3	61.4	32.9	19.7	11.2	31.2
Attention [2]	28.1	26.8	57.3	40.1	29.3	25.1	34.5
Proposal	92.3	76.8	93.1	93.7	70.8	86.2	**85.5**
	Fashion-MNIST						
Traditional	21.3	24.6	36.9	28.5	29.6	16.8	26.7
Attention [2]	18.2	20.1	51.8	26.0	31.8	36.2	30.7
Proposal	62.9	61.5	66.5	60.9	60.8	45.9	**59.8**

Table 2: Effect of varying the random noise rate in terms of Accuracy (%).

	0.0	0.01	0.05	0.1	0.2	0.3	0.4	0.5
Baseline	90.1	33.3	13.2	11.2	10.5	10.2	10.2	10.6
Proposal	70.7	70.1	61.6	56.7	45.5	43.2	39.6	35.0

Table 3: Average accuracy for Fashion-MNIST when using different mask or background generators.

	Proposal	True mask	Noise background
Accuracy (%)	59.8	72.8	10.0

For that purpose, an adversarially trained model is proposed that is invariant to the background. During training, the target model tries to minimize its loss, but a generator counteracts it by injecting new backgrounds by optimizing for backgrounds that maximize the loss, thus making the target model robust to background changes. The method is evaluated using a synthetic dataset.

While the proposed method was evaluated for the task of classification, it could potentially be used for other tasks involving a CNN, such as regression problems, segmentation, or reinforcement learning tasks.

Acknowledgments

This work is financed by the ERDF – European Regional Development Fund through the Operational Programme for Competitiveness and Internationalisation – COMPETE 2020 Programme and by National Funds through the Portuguese funding agency, FCT - Fundação para a Ciência e a Tecnologia within project "POCI-01-0145-FEDER-028857". Furthermore, Ricardo Cruz was supported by Ph.D. grant SFRH/BD/122248/2016, also provided by FCT.

References

[1] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[2] Qingji Guan, Yaping Huang, Zhun Zhong, Zhedong Zheng, Liang Zheng, and Yi Yang. Diagnose like a radiologist: Attention guided convolutional neural network for thorax disease classification. *arXiv*, 2018.

[3] Yann LeCun, Corinna Cortes, and CJ Burges. MNIST handwritten digit database. *AT&T Labs [Online]. Available: http://yann.lecun.com/exdb/mnist*, 2010.

[4] Ricardo M. Prates, Ricardo Cruz, AndrÃl' P. Marotta, Rodrigo P. Ramos, Eduardo F. Simas Filho, and Jaime S. Cardoso. Insulator visual non-conformity detection in overhead power distribution lines using deep learning. *Computer and Electrical Engineering*, 2019. doi: 10.1016/j.compeleceng.2019.08.001.

[5] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[6] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms. *CoRR*, 2017.

Semantic *Vs* Radiomic Features from CT Images to Predict Gene Mutation Status in Lung Cancer

Tania Pereira[1]
tania.pereira@inesctec.pt
Gil Pinheiro[1]

Catarina Dias[12]

António Cunha[13]
acunha@utad.pt
Hélder P. Oliveira[14]
helder.f.oliveira@inesctec.pt

[1] INESC TEC - Institute for Systems and Computer Engineering, Technology and Science, Portugal
[2] FEUP - Faculty of Engineering, University of Porto, Porto, Portugal
[3] UTAD - University of Trás-os-Montes and Alto Douro, Vila Real, Portugal
[4] FCUP - Faculty of Science, University of Porto, Porto, Portugal

Abstract

In lung cancer, the biopsy is the traditional method to assess the mutation status of the most frequent and relevant oncogenes. Medical imaging, which is already a common source of information in clinical practice, is a potential alternative to the biopsy. It contains a large number of features that, although not visible to the naked eye, may be valuable for tumour characterisation. The recent field of radiomics allows new opportunities for the genomic analysis of a tumour, by extracting hundreds of quantitative features from medical images which, in a non-invasive way, provide a full state visualisation of a tumour at a macroscopic level. This study aimed to investigate in which extent features extracted from medical images are related to helpful genotype factors for tumour characterisation, in particular for *EGFR* and *KRAS* mutation status. Radiomic and semantic features were used for the prediction. The performance of the models demonstrated that *EGFR* (AUC=0.75) mutation status can be differentiated through medical images using semantic features. The experiments suggest that the best way to approach this problem is by combining nodule-related features with features from other lung structures.

1 Introduction

Lung cancer is the cancer type leading the incidence and mortality rates [5]. This is linked to the fact that it is often diagnosed in an advanced stage, which magnifies the importance of treatments for advanced-stage disease. Epidermal Growth Factor Receptor (*EGFR*) and Kristen Rat Sarcoma Viral Oncogene Homolog (*KRAS*) are the most frequently mutated gene in lung cancer [8]. Current molecularly-targeted therapies can effectively target specific biomarkers, decreasing multiple undesirable side effects associated with cancer treatment. Radiogenomics, a specific field within radiomics, is defined by the correlation between quantitative features, directly extracted from radiological images (imaging phenotype), and genetic information (genotype). Studies in lung cancer have presented the association between *EGFR* mutation status and quantitative features extracted from computed tomography (CT) scans [1, 4]. This study aims to provide further advances and to open new discussions in the lung cancer radiogenomics field by exploring the data and building machine learning models, while considering different subsets of inputs. More specifically, predictive models for *EGFR* and *KRAS* mutation status in lung cancer were developed. The current paper is an adaptation of our previously published work [9].

2 Material and Methods

2.1 Dataset

The NSCLC-Radiogenomics dataset [7] comprises data collected between 2008 and 2012 from a cohort of 211 patients with Non-small-cell lung cancer (NSCLC) referred for surgical treatment, being the only public dataset which comprehends information regarding the mutation status of lung cancer-related genes (*EGFR* and *KRAS*). It contains a set of CT images stored in DICOM format.

2.1.1 Molecular Data

Despite including a cohort of 211 NSCLC subjects, only 116 (wild type: 93, mutant: 23) were further considered in the presented radiomic study for *EGFR* mutation status prediction and 114 (wild type: 88, mutant: 26) for *KRAS* mutation status prediction. The scarce availability of tumour masks and target labels did not allow all subjects to be used.

2.1.2 Clinical Features

Clinical features were added to the radiomic features as well as to the semantic features to build the predictive models.

2.1.3 Radiomic Features

There are image properties, such as the distance between slices, which may differ from scan to scan, and consequently affect the features extracted and the learning ability of the algorithms. Therefore, before trying to extract patterns, the images went through a preprocessing step in order to standardise the scans across the whole dataset. The CT image values were converted to Hounsfield Unites (HU), which is a measure of radiodensity. From the 3D images of the nodules of the pre-processed CT scans, a set of 1218 radiomic features were extracted using the open-source package *Pyradiomics* [10]. Features were computed both on the original image and on images obtained after application of wavelet and Laplacian of Gaussian (LoG) filters. Six classes of features were extracted from the *Pyradiomics* package: shape-based features (14 features), first-order features (18 features), GLCM features (22 features), GLRLM features (16 features), GLSZM features (16 features) and GLDM features (14 features).

2.1.4 Semantic Features

The dataset comprises a set of subjects whose tumour was analysed by radiologists using 30 nodule and parenchymal features, which describe nodule's geometry, location, internal features and other related findings. From these subjects, 158 are characterised in terms of *EGFR* mutation status and 157 subjects characterised in terms of *KRAS* mutation status, which were the samples selected for the presented semantic study.

2.2 Balancing Training Set

In general, machine learning algorithms assume a similar distribution of classes. *EGFR* wild type is over-represented, which could result in a model biased towards this class. To overcome this class imbalance, Synthetic Minority Over-sampling Technique - Nominal and Continuous (SMOTE-NC) was applied, an extended version of SMOTE generalised to handle data with continuous and nominal features [2].

2.3 Classification and Feature Importance

The classifier used in this work was Extreme Gradient Boosting (XGBoost), which is a scalable and accurate implementation of gradient boosted trees algorithms [3]. A benefit of using gradient boosting is that after the boosted trees are constructed, it is possible to retrieve the importance scores for each feature, based on how useful or valuable each feature was in the construction of the boosted decision trees within the model.

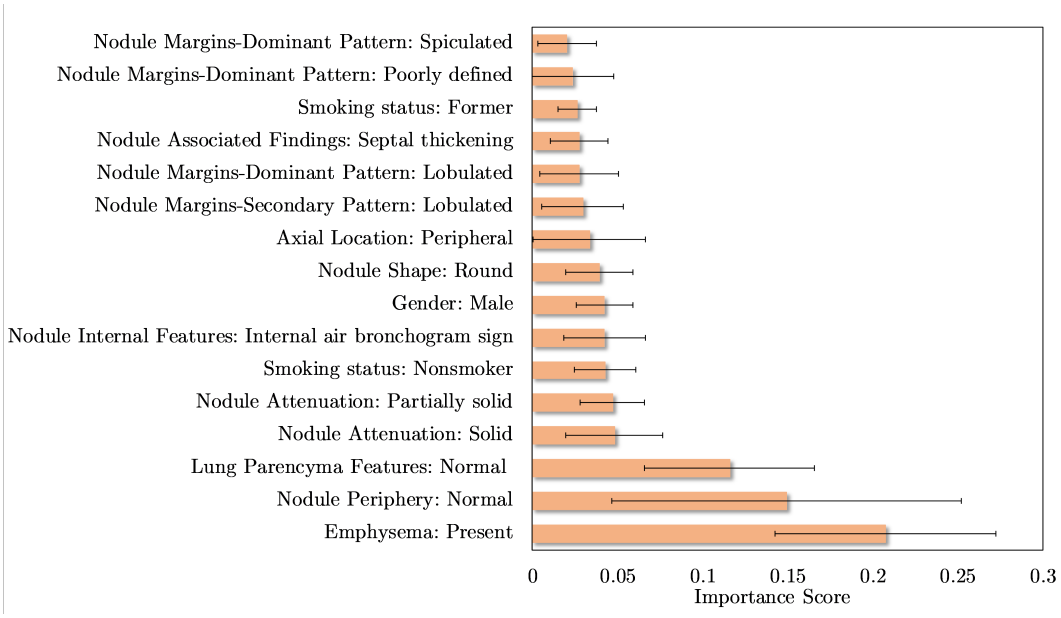


Figure 1: Top 16 semantic features based on the importance scores of features, measured via XGBoost, for predicting the *EGFR* mutation status.

3 Results

Mean values of Area Under the Curve (AUC) were reported for 100 random data splits, with a division of 80% and 20% for training and testing, respectively. Two main types of input features were considered: radiomic and semantic. The semantic were further divided into features that only describe the nodule, features that only describe structures external to nodule and a hybrid between the previous two. Radiomics were not further divided as they only describe the nodule. We designed those four experiments in order to test and compare which type of input features allow to achieve better performance in gene mutation status prediction (Table 1). Only the predictive models for *EGFR* showed relevant results, with a maximum mean AUC of 0.7458 ± 0.0877 using the hybrid semantic features (Table 1). A subset of features, ranked by importance for the most successful model (*EGFR* mutation status prediction using hybrid semantic features), is presented in Figure 1. They were selected using a minimum threshold of 0.02 and add up to cumulative importance of 0.92 out of 1.

AUC	*EGFR* Prediction	*KRAS* Prediction
Radiomic	0.5797 ± 0.1238	0.5087 ± 0.0104
Semantic Nodule	0.6542 ± 0.0953	0.4381 ± 0.0679
Semantic Non-Nodule	0.6831 ± 0.0890	0.4921 ± 0.0851
Semantic Hybrid	0.7458 ± 0.0877	0.5035 ± 0.0776

Table 1: Classification results for *EGFR* and *KRAS* mutation status predictive models.

4 Discussion and Conclusions

The results of the present study suggest that even though *EGFR* mutation status is correlated to CT scans imaging phenotypes, the same does not hold true for *KRAS* mutation status. We hypothesise that this might be due to two reasons: mutated and wild type *KRAS* display identical imaging phenotypes, which is supported by the literature [6, 11], or our number of samples was too small and unrepresentative to find a relevant pattern for such a complex problem. The outcomes of this work also indicate that general lung semantic features in conjunction with tumour specific semantic features should be used in order to obtain the best possible *EGFR* mutation status classification results. This, combined with the fact that the most relevant features (as determined by the classifier) were tumour external, might hint towards the importance of a holistic lung analysis, instead of a local nodule analysis. It is crucial to emphasise this characteristic, as it might change the direction and broaden the analysis spectrum of future radiogenomics studies, which until now have been mainly focusing on the nodule or in a region of interest around it [12]. Since there is a large spectrum of clinicopathological processes that occur during the lung cancer development, it is only natural that important information for the predictive models can be obtained from a larger region of analysis that contains other structures from the lung.

Acknowledgment

This work is financed by National Funds through the Portuguese funding agency, FCT within project UIDB/50014/2020.

References

[1] Z. Bodalal, S. Trebeschi, T. D. L. Nguyen-Kim, W. Schats, and R. Beets-Tan. Radiogenomics: bridging imaging and genomics. *Abdominal Radiology*, 2019.

[2] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

[3] T. Chen and C. Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016.

[4] S. R. Digumarthy, A. M. Padole, R. L. Gullo, L. V. Sequist, and M. K. Kalra. Can ct radiomic analysis in nsclc predict histology and egfr mutation status? *Medicine*, 98(1), 2019.

[5] J. Ferlay, I. Soerjomataram, R. Dikshit, et al. Cancer incidence and mortality worldwide: Sources, methods and major patterns in GLOBOCAN 2012. *International Journal of Cancer*, 2015.

[6] O. Gevaert, S. Echegaray, A. Khuong, et al. Predictive radiogenomics modeling of egfr mutation status in lung cancer. *Scientific reports*, 7:41674, 2017.

[7] O. Gevaert, J. Xu, C. D. Hoang, et al. Non-small cell lung cancer: Identifying prognostic imaging biomarkers by leveraging public gene expression microarray data - Methods and preliminary results. *Radiology*, 2012.

[8] S. E. Jorge, S. S. Kobayashi, and D. B. Costa. Epidermal growth factor receptor (EGFR) mutations in lung cancer: Preclinical and clinical data, 2014.

[9] T. Pereira, G. Pinheiro, C. Dias, et al. Identifying relationships between imaging phenotypes and lung cancer-related mutation status: EGFR and KRAS. *Scientific Reports*, 10:3625, 2020.

[10] J. J. Van Griethuysen, A. Fedorov, C. Parmar, et al. Computational radiomics system to decode the radiographic phenotype. *Cancer research*, 77(21):e104–e107, 2017.

[11] S. S. Yip, J. Kim, T. P. Coroller, et al. Associations between somatic mutations and metabolic imaging phenotypes in non-small cell lung cancer. *Journal of Nuclear Medicine*, 2017.

[12] W. Zhao, J. Yang, B. Ni, et al. Toward automatic prediction of EGFR mutation status in pulmonary adenocarcinoma with 3D deep learning. *Cancer Medicine*, 2019.

Active Robot Learning for Efficient Body-Schema Online Adaptation

Gonçalo Cunha
goncalocarvalhocunha@gmail.com
Alexandre Bernardino
alexandre.bernardino@tecnico.ulisboa.pt
Pedro Vicente
pvicente@isr.tecnico.ulisboa.pt
Ricardo Ribeiro
ribeiro@isr.tecnico.ulisboa.pt
Plínio Moreno
plinio.lopez@tecnico.ulisboa.pt

Instituto de Sistemas e Robótica
Universidade de Lisboa - Instituto Superior Técnico
Lisboa, Portugal

Abstract

This work proposes an active learning approach for estimating the Denavit-Hartenberg parameters of 7 joints of the iCub arm in a simulation environment, using observations of the end-effector's pose and knowing the values from proprioceptive sensors. Cost-sensitive active learning, aims to reduce the number of measurements taken and also reduce the total movement performed by the robot while calibrating, thus reducing energy consumption, along with mechanical fatigue and wear. The estimation of the arm's parameters is done using the Extended Kalman Filter and the active exploration is guided by the A-Optimality criterion. The results show cost-sensitive active learning can perform similarly to the straightforward active learning approach, while reducing significantly the necessary movement.

1 Introduction

Active learning is a sub-field of machine learning which aims to reduce the amount of training data required to build a model, with a certain precision. This is done by having the learning algorithm decide which data it wants to label/sample next. A general introduction for this area of research can be found in [8].

1.1 Related Work

Recent works have succeeded in employing different strategies for body schema adaptation, such as [9], [10] and [7]. All these works show different successful ways of accounting for the robot's body errors, but using active learning to this effect would promote faster adaptation.

Active learning methods have better empirical results, when compared to random sampling. Some of these works are described in [3], [4], [7], and [2]. These works have shown the advantages of using active learning but assume all samples have equal acquisition cost. In [5] a criterion is proposed considering uncertainty and travel cost for the designated task, minimising the accumulated path length needed for accurate estimation, at the cost of an increase of the number of samples.

1.2 Contributions

This work aims to estimate the Denavit-Hartenberg (DH) parameters of 7 rotational joints of the iCub arm in a simulation environment, by acquiring observations from the pose of the end-effector, using active learning to select the best joint configurations for movement and sampling efficiency.

By portraying an arbitrary serial robotic arm as in Figure 1, a calibration routine is proposed to make use of active learning to select the best joint configurations to sample the end-effector pose, in order to estimate the DH parameters with the best possible precision, using the Extended Kalman Filter (EKF).

Similarly to [5], we argue that using active learning to reduce the number of samples taken may not be the best approach, since some of the best samples may require unnecessary long movements, increasing execution time and energy spent. The cost-sensitive active learning approach provides a tunable trade-off between minimising the number of iterations required and minimising the required movement. The proposed calibration routine is composed of the key steps shown in Figure 2.

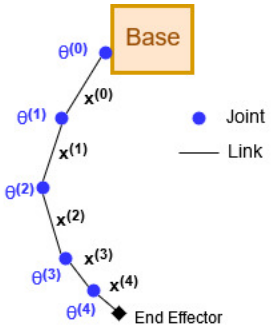


Figure 1: Illustration of a robot's kinematic chain. $\theta^{(i)}$ represents the angle value for joint i and $x^{(i)}$ represents the Denavit-Hartenberg parameters describing the transformation between frames i and $i+1$.

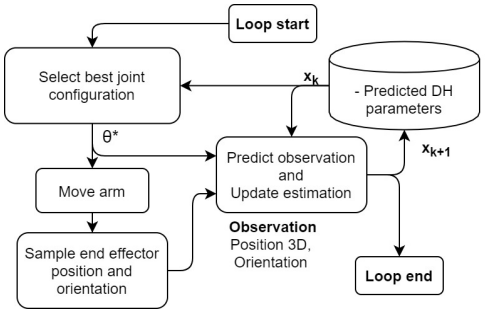


Figure 2: Key steps in the structure of the required program.

1.3 Extended Kalman Filter

The EKF, explained in detail in [6], allows recursive parameter estimation of systems represented by a nonlinear model, which is the case for the relation between the DH parameters, x, and the end-effector pose, z, given by the function

$$z = h(x, \theta), \tag{1}$$

where θ represents the known joint angles. Since the DH parameters are constant in time, the EKF can be summarized in the following 3 equations. The predicted co-variance, P, of the DH parameters, x, is given by

$$P_{k+1|k} = P_{k|k} + Q_k, \tag{2}$$

where Q_k is the co-variance matrix of the Gaussian noise associated with slow changes in the DH parameters, e.g. due to temperature. The update of the prediction, $\hat{x}$, after obtaining a measurement, z_k, is given by

$$\hat{x}_{k+1|k+1} = \hat{x}_{k+1|k} + K_{k+1}[z_k - h(\hat{x}_{k+1|k}, \theta_k)] \tag{3}$$

and the update of the co-variance, P, is given by

$$P_{k+1|k+1} = P_{k+1|k} - K_{k+1}\left[H_k P_{k+1|k} H_k^T + R_{k+1}\right] K_{k+1}^T, \tag{4}$$

where

$$K_{k+1} = P_{k+1|k} H_{k+1}^T \left[H_k P_{k+1|k} H_k^T + R_{k+1}\right]^{-1}, \tag{5}$$

H is the jacobian matrix of the observation function in (1), with respect to x, $\frac{\partial h}{\partial x}$, and R is the co-variance matrix of the Gaussian noise present in the measurements.

Link	0	1	2	3	4	5	6
a [mm]	0	0	-15	15	0	0	62.5
d [mm]	-107.74	0	-152.28	0	-137.4	0	16
α [rad]	$\frac{\pi}{2}$	$-\frac{\pi}{2}$	$-\frac{\pi}{2}$	$\frac{\pi}{2}$	$\frac{\pi}{2}$	$\frac{\pi}{2}$	0
θ [rad]	$-\frac{\pi}{2}$	$-\frac{\pi}{2}$	$-\frac{7\pi}{12}$	0	$-\frac{\pi}{2}$	$\frac{\pi}{2}$	π

Table 1: Actual DH parameters of the iCub arm in the iCub simulator.

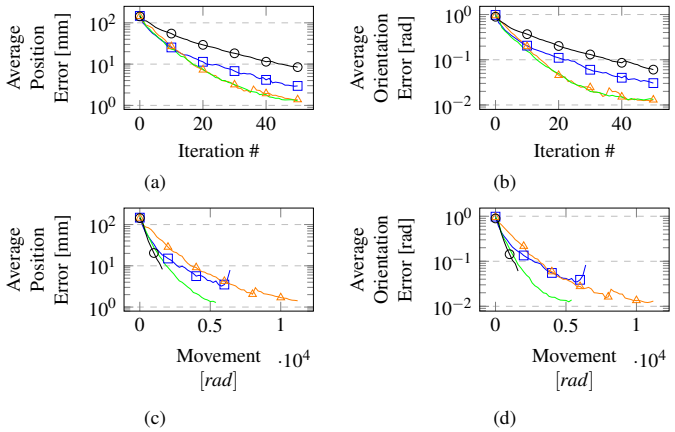


Figure 3: Mean error evolution while performing the calibration routine for different values of δ: (a) Average position error evolution in millimetres at each iteration; (b) Average orientation error evolution in radians at each iteration; (c) Average position error evolution in millimetres with respect to joint movement; (d) Average orientation error evolution in radians with respect to joint movement. Legend: Square - Random; Triangle - $\delta = 1$; No marker - $\delta = 0.4$; Circle - $\delta = 0.1$.

1.4 Cost-sensitive Active Learning

This work aims to choose the best joint configurations to sample the end-effector pose, at each iteration of the calibration routine, to reduce both the body-schema error and movement performed while calibrating. Martinez-Cantin *et al.* [2] successfully used the A-optimality criterion to reduce the number of samples taken. It consists in choosing the joint angles θ which minimise the expected mean squared error of the robot parameters, x, which approximates to minimising the expected trace of the co-variance matrix, P. As described, the cost function is given by $C(\theta) = \mathbb{E}\left[(\hat{x}_{k+1} - x)^T(\hat{x}_{k+1} - x)|z_{1:k}, \theta_{1:k}\right] \approx \mathbb{E}\left[tr(P_{k+1})|z_{1:k}, \theta_{1:k}\right]$.

This work proposes adding constraints to the optimisation problem as in

$$\theta_k^* = \underset{\theta \in \left[\theta_{k-1}^* - \Delta, \theta_{k-1}^* + \Delta\right]}{\operatorname{argmin}} \quad C(\theta), \tag{6}$$

where θ_{k-1} is the previous joint configuration selected and Δ is a vector of size n (number of joints), defining the boundaries of the search space. Considering normalised joint values in the interval $[0, 1]$, Δ is defined as $\Delta = \delta \cdot \mathbf{1_n}$, where $\mathbf{1_n}$ is a unit vector of size n and δ is a tuning parameter that defines the relative movement every joint can perform around the current arm configuration. Since the problem defined in (6) is not convex, it must be solved using a global optimisation method, for which it is used the DIRECT algorithm, proposed in [1].

2 Results

Results were obtained for different values of δ. For each different value, the calibration routine runs fifty times and the plots from Figure 3 show the average values of position and orientation error. At each run, the DH parameters are initialised with values from a uniform distribution, where the means are the actual values of the DH parameters, from Table 1 and the width of the distribution is 30% of the highest value from all the linear and angular DH parameters, 46 mm and 0.94 rad, respectively. The position error is given by the euclidean distance between the predicted position and actual position of the end-effector and the orientation error is given by computing $d(R_A, R_B) = \sqrt{\frac{\|logm(R_A^T R_B)\|_F^2}{2}}$ [rad], between the predicted, R_A, and actual, R_B, end-effector rotation matrices, where $logm$ is the principal matrix logarithm and $\|\cdot\|_F$ is the Frobenius norm. Gaussian noise is added to the observations with a standard deviation of 2 mm for the position coordinates and 0.08 radians for the orientation.

Looking at Figures 3(a) and 3(b), the advantages of using the active learning method proposed in [2], corresponding to $\delta = 1$, can be observed by comparing it with selecting random joint configurations to sample, instead of solving (6), since there is a more significant reduction in error at each iteration. In Figures 3(c) and 3(d), the same data is represented, but the x axis represents the movement performed by the arm. It is visible the amount of extra movement performed by the active learning method, $\delta = 1$, almost double of the random method. Restricting movement, making $\delta = 0.4$, yields no performance loss, regarding Figures 3(a) and 3(b), and it is more efficient, as Figures 3(c) and 3(d) show.

3 Conclusions

The results show there is an advantage in restricting movement during the optimisation stage. It is possible to reduce the movement performed by roughly half and still maintain the iteration wise performance. If movement efficiency is a priority, one can restrict the movement even more, at the cost of more iterations. It is worth mentioning, more iterations does not mean lower time-efficiency, since reducing the amount of time spent moving may make up for the extra computing time. Indeed, it will depend on the computing power and the speed at which the arm moves.

For future work, it is planned to obtain results using the iCub cameras and fiducial markers placed on its hand. This comes with observation noise dependant on the observed pose and it should be taken into account when selecting optimal joint configurations.

Acknowledgements

This work was supported by FCT with the LARSyS - FCT Project UIDB/50009/2020 and the PhD grant PD/BD/135115/2017.

References

[1] D. R. Jones, C. D. Perttunen, and B. E. Stuckman. Lipschitzian optimization without the Lipschitz constant. *Journal of Optimization Theory and Applications*, 79(1):157–181, 1993.

[2] Ruben Martinez-Cantin, Manuel Lopes, and Luis Montesano. Body schema acquisition through active learning. *Proceedings - IEEE International Conference on Robotics and Automation*, pages 1860–1866, 2010.

[3] Uriel Martinez-Hernandez, Tony J. Dodd, Mathew H. Evans, Tony J. Prescott, and Nathan F. Lepora. Active sensorimotor control for tactile exploration. *Robotics and Autonomous Systems*, 87:15–27, 2017.

[4] Uriel Martinez-Hernandez, Tony J. Dodd, and Tony J. Prescott. Feeling the Shape: Active Exploration Behaviors for Object Recognition with a Robotic Hand. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 48(12):2339–2348, 2018.

[5] Takamitsu Matsubara and Kotaro Shibata. Active tactile exploration with uncertainty and travel cost for fast shape estimation of unknown objects. *Robotics and Autonomous Systems*, 91:314–326, 2017.

[6] Nancy Reid. *Estimation. In: Lovric M. (eds) International Encyclopedia of Statistical Science*, pages 455–459. Springer, Berlin, Heidelberg, 2011.

[7] Arturo Ribes, Jesus Cerquides, Yiannis Demiris, and Ramon Lopez de Mantaras. Active Learning of Object and Body Models with Time Constraints on a Humanoid Robot. *IEEE Transactions on Cognitive and Developmental Systems*, 8(1):26–41, 2015.

[8] Burr Settles. Active Learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 6(1):1–114, 6 2012.

[9] Pedro Vicente, Lorenzo Jamone, and Alexandre Bernardino. Online body schema adaptation based on internal mental simulation and multisensory feedback. *Frontiers Robotics AI*, 3(MAR), 2016.

[10] Rodrigo Zenha, Pedro Vicente, Lorenzo Jamone, and Alexandre Bernardino. Incremental adaptation of a robot body schema based on touch events. *2018 Joint IEEE 8th International Conference on Development and Learning and Epigenetic Robotics, ICDL-EpiRob 2018*, pages 119–124, 2018.

A Deep Image Segmentation Approach to Breast Keypoint Detection

Tiago Gonçalves [1,2]
tiago.f.goncalves@inesctec.pt
Wilson Silva [1,2]
wilson.j.silva@inesctec.pt
Jaime S. Cardoso [1,2]
jaime.cardoso@inesctec.pt

[1] Faculdade de Engenharia
Universidade do Porto
Porto, Portugal

[2] INESC TEC
Porto, Portugal

Abstract

The main aim of breast cancer conservative treatment is the optimisation of the aesthetic outcome and women's quality of life, without jeopardising local cancer control and overall survival. Recently, a deep learning algorithm, used in conjunction with a shortest-path algorithm that models images as graphs, has been proposed and achieved state-of-the-art results. However, it is both time-consuming and computationally complex. In this work, we studied a novel algorithm based on the interaction of deep image segmentation and deep keypoint detection models capable of improving both performance and execution-time on the breast keypoint detection task.

1 Introduction

Breast cancer ranks as the most frequent cancer among women [1, 5]. Despite being a highly mutable and rapidly evolving disease, it is estimated that most breast cancers are curable if properly detected and treated [8]. Under this paradigm, it is possible to surgically treat most cancers with a breast cancer conservative treatment (BCCT), which does not require the removal of the entire breast, as in mastectomy [12]. Currently, to perform the aesthetic assessment of BCCT, the majority of the extracted features are related to asymmetry measurements [4]. To facilitate the extraction of such features, it is fundamental to mark breast keypoints. The advent of machine learning and deep learning opened the possibility to design novel algorithms based on deep neural networks (DNN) which may be fully end-to-end (*i.e.*, receive an image and output the aesthetic assessment score). Until the publication of this work [6], the state-of-the-art algorithm for keypoint detection was a hybrid method based on a DNN and on traditional computer vision methods, which made it computationally heavy and slow. This work presents the development of a novel breast keypoint detection algorithm that addresses the efficiency problem while maintaining or improving the accuracy. A study of algorithm performance based on execution time has also been conducted, since, at the long term, the intention is to deploy such algorithms into a web-based application that could be accessed by a diversity of devices and operative systems.

2 Deep Keypoint Detection Algorithm

Silva *et al.* proposed a method [10] that uses a deep neural network (DNN) for the keypoint detection task, opening the possibility to follow an integrated learning approach. Following the ideas of Cao *et al.* [3] and Belagiannis *et al.* [2], Silva *et al.* proposed an architecture that first learns how to regress *heatmaps* (in which the ground-truth was obtained by applying a Gaussian kernel to the keypoints) and, after iterative refinement of this heatmap regression, it can predict keypoint localisation. The heatmap regression is performed with the U-Net model [9]. Then, to do the keypoint regression, the original images are multiplied by the refined heatmaps (to improve the initial fuzzy localization of keypoints) and are fed to a keypoint regression module composed of three blocks: top layers of the VGG16 [11] (*i.e.*, only the convolutional layers), four additional convolutional layers and three dense layers. The entire model is trained, using the iterative refinement of the regression of heatmaps as a regularisation term of the loss function. The endpoints and the nipples are obtained with this deep learning algorithm, whereas the contour is refined with the shortest-path algorithm, which models images as graphs, based on gradients (see Figure 1). However, this procedure is very time-consuming when compared with the inference process of a deep learning model only. Also, if one intends to integrate such algorithms into a web application, it is of utmost importance that performance measurements (*e.g.*, loading time, execution time) are taken into account when testing and designing novel methods.

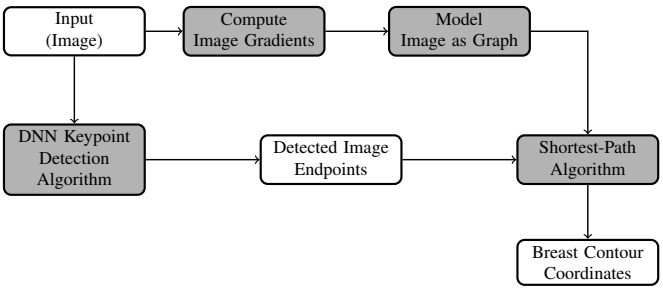


Figure 1: Hybrid keypoint detection algorithm, from [10].

3 Deep Image Segmentation for Breast Keypoint Detection

The intuition behind this approach is that it is easier to detect breast contours if one is capable to detect breasts first [7]. This can be seen as a problem of semantic segmentation, where both breasts are considered the foreground and the rest is considered the background of the image. The main hypothesis is that if it is possible to perform the segmentation of both breasts with high precision, one could proceed to an algorithm of contour detection and then accurately extract the keypoints related to the breast contours. With segmentation, the goal is to learn a single solution (*i.e.*, one image corresponds to one mask). This is important because, if the DNN is capable of predicting the correct mask, the set of points of the detected contour will contain a subset of points that belong to the real breast contour. On the other hand, with keypoint regression, there is a higher degree of variability, where the DNN can predict points that belong to the real breast contour and points that do not, negatively influencing the algorithm performance. Furthermore, when compared with the hybrid keypoint detection algorithm, one expects that this approach will bring improvements in terms of results and performance, *i.e.*, it would be faster than the hybrid keypoint detection algorithm.

4 Implementation and Results

The available dataset (221 images) has 37 ground-truth keypoints (4 endpoints, 30 points along the breast contours, 2 nipples and the sternal notch) resulting in a total of 74 coordinates per image. All experiences were done taking into account 5-fold cross-validation split into train and test sets. First, we trained a deep image segmentation model that could achieve good results in breast segmentation. To train this model, it was necessary to generate ground-truth breast masks, which were obtained with the support of the ground-truth keypoints and images.

Taking into account previous results with other models [7], for this experiment, we decided to use the U-Net++ [13] architecture as the deep segmentation model. The U-Net++ model was trained and used to generate segmentation masks. From these masks, contours were extracted. This first step outputs a variable number of contour keypoints, some of which are not desired, since they do not belong to what is considered the breast contour. As a post-processing step, we project the Silva *et al.* DNN predicted keypoints onto to the mask contours through the minimization of the Euclidean Distance between the mask contour keypoint and the predicted keypoint (see Figure 2). At the end of this processing step, the final set of breast keypoints is obtained (see Figure 3). We also studied the

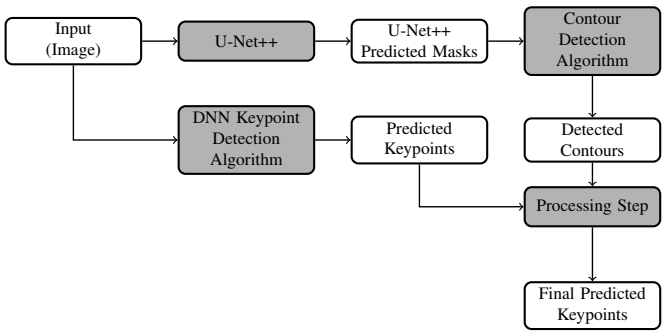


Figure 2: Scheme of the deep image segmentation algorithm for breast keypoint detection.

performance of both ours and Silva *et al.* algorithms to assess which one would fit better into a web-version of BCCT.core, capable of real-time interaction with deep learning models. To perform this study, the execution time of each algorithm on CPU (Intel® Core™ i7-2600 CPU @ 3.40GHz × 8) was measured on the test set of each cross-validation fold. Table 1 presents the average error distance (measured in pixels) and the average execution time (measured in seconds) of each model inference on the test set. It can be seen that our proposed method surpasses both the DNN and hybrid keypoint detection algorithms from Silva *et al.* in the endpoints and breast contours detection tasks, which were the state-of-the-art breast keypoint detection algorithms. Moreover, this novel algorithm achieves lower values of standard deviation and maximum error, which suggests it is even more robust when compared with the other two. Regarding the study of performance, it can be understood that the DNN keypoint detection algorithm achieves better execution time, however, it has the highest error for the breast contour. Our method presents the best balance between time-efficiency and accuracy, as it is the most accurate model, and has a time efficiency comparable to the most time-efficient method.

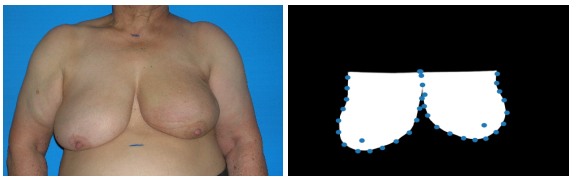


Figure 3: Example of the results obtained with the proposed deep image segmentation method. The first image is the input photograph and the second image is the U-Net++ predicted mask with the detected breast keypoints (after the processing step).

5 Conclusions and Future Work Recommendations

In this work we presented a novel algorithm based on the interaction of two deep learning models (a segmentation model, and a keypoint detection model) that can surpass the state-of-the-art algorithms present in the literature [6]. Furthermore, a comparative study regarding algorithms performance was done to assess which one would fit better a web-based application for the aesthetic assessment of BCCT. Our proposed model revealed itself as the best in terms of keypoint prediction, while being very competitive in terms of executing time. As future work, the next step will be to improve results on nipples detection task and to modify this novel segmentation-based keypoint detection algorithm by integrating all the tasks of its pipeline in a unique DNN with a combined loss function. The integration and full deployment of this algorithm in a web-application are also planned.

Table 1: Average error distance for endpoints, breast contours and nipples, measured in pixels and average execution time of the models' inferences. Best results are highlighted in bold. Note: STD stands for standard deviation and Max stands for maximum error.

Model	Endpoints			Breast Contours			Nipples			Execution Time (s)
	Mean	STD	Max	Mean	STD	Max	Mean	STD	Max	
DNN keypoint detection algorithm	40	**33**	**182**	21	8	72	**70**	**39**	**218**	**150**
Hybrid keypoint detection algorithm	40	**33**	**182**	13	14	104	**70**	**39**	**218**	1704
Our keypoint ketection algorithm	**38**	34	195	**11**	**5**	**34**	**70**	**39**	**218**	280

Acknowledgements

The project "TAMI - Transparent Artificial Medical Intelligence" (NORTE-01-0247-FEDER-045905) leading to this work is co-financed by ERDF - European Regional Fund through the Operational Program for Competitiveness and Internationalisation - COMPETE 2020, the North Portugal Regional Operational Program - NORTE 2020 and by the Portuguese Foundation for Science and Technology - FCT under the CMU - Portugal International Partnership and the PhD grants "SFRH/BD/139468/2018" and "SFRH/BD/06434/2020".

References

[1] *World cancer report 2014*. ISBN 978-92-832-0443-5.

[2] Vasileios Belagiannis and Andrew Zisserman. Recurrent Human Pose Estimation. *arXiv:1605.02914 [cs]*, May 2016. arXiv: 1605.02914.

[3] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime Multi-Person 2d Pose Estimation using Part Affinity Fields. *arXiv:1611.08050 [cs]*, November 2016. arXiv: 1611.08050.

[4] Jaime S. Cardoso and Maria J. Cardoso. Towards an intelligent medical system for the aesthetic evaluation of breast cancer conservative treatment. *Artificial Intelligence in Medicine*, 40(2):115–126, June 2007. ISSN 09333657. doi: 10.1016/j.artmed.2007.02.007.

[5] Sharon H Giordano, Aman U Buzdar, and Gabriel N Hortobagyi. Breast Cancer in Men. page 11.

[6] Tiago Gonçalves, Wilson Silva, Maria J Cardoso, and Jaime S Cardoso. A novel approach to keypoint detection for the aesthetic evaluation of breast cancer surgery outcomes. *Health and Technology*.

[7] Tiago Gonçalves, Wilson Silva, and Jaime Cardoso. Deep Aesthetic Assessment of Breast Cancer Surgery Outcomes. Springer International Publishing, Cham, 2020. doi: 10.1007/978-3-030-31635-8_236.

[8] Helder P. Oliveira, Jaime S. Cardoso, Andre Magalhaes, and Maria J. Cardoso. Methods for the Aesthetic Evaluation of Breast Cancer Conservation Treatment: A Technological Review. *Current Medical Imaging Reviews*, 9(1):32–46, April 2013. ISSN 15734056. doi: 10.2174/1573405611309010006.

[9] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. *arXiv:1505.04597 [cs]*, May 2015. arXiv: 1505.04597.

[10] Wilson Silva, Eduardo Castro, Maria J. Cardoso, Florian Fitzal, and Jaime S. Cardoso. Deep Keypoint Detection for the Aesthetic Evaluation of Breast Cancer Surgery Outcomes. In *Proceedings of the IEEE International Symposium on Biomedical Imaging (ISBI'19)*, 2019.

[11] Karen Simonyan and Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv:1409.1556 [cs]*, September 2014. URL `http://arxiv.org/abs/1409.1556`. arXiv: 1409.1556.

[12] Williams Street. Cancer Facts & Figures 2018. page 76, 2018.

[13] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. UNet++: A Nested U-Net Architecture for Medical Image Segmentation. Springer International Publishing, Cham, 2018. doi: 10.1007/978-3-030-00889-5_1.

Fire and Smoke recognition in crowdsourced images with YOLO networks

Ana Madeira
abreumadeira.ana@gmail.com
Catarina Silva
catarina@dei.uc.pt
Alberto Cardoso
alberto@dei.uc.pt
Bernardete Ribeiro
bribeiro@dei.uc.pt

University of Coimbra
CISUC - Centro de Informática e Sistemas
FCTUC-DEI - Departamento de Engenharia Informática
Coimbra, Portugal

Abstract

The early detection of a fire can largely mitigate harmful consequences. With the improvement in image quality, it is now possible to develop intelligent systems for visually detecting forest fires. An intelligent system for fire detection was implemented based on deep learning techniques for image object detection. As part of the fire detection approach development, different datasets are proposed to train and evaluate the YOLO models, specific to the fire and smoke recognition problem. The proposed Fire/Smoke annotated datasets can be used in future smoke, and fire detection research. Results show that a YOLOv4 one-stage detector can be used for image fire and smoke detection tasks, trained using manually annotated datasets and applied to a real application using crowdsourced data.

1 Introduction

As a way for people to report fires detected using their smartphones, the FireLoc project [1] of the Foundation of Science and Technology (FCT) [2] is set up as an alternative way of reporting fires. This project is based on voluntary contributions and aims to develop a system in which, through a smartphone application, users can send photos of fire taken with their smartphone camera. If present, smoke and fire are recognized in the images submitted, and the forest fire can be located on a map. This information is then sent to a server. The developed system will correspond to the submission validation module and validate whether there is fire or smoke in each contribution.

In this paper, the main focus is the development of an intelligent system for fire and smoke detection. integrated in the FireLoc application, using the proposed post-processing steps to obtain the image classification results, identifying whether user submissions are valid, i.e., whether they contain smoke or fire.

2 Related work

Recent studies show advantages in considering the localization as well as the classification of existing objects in an image as part of object detection problems [7]. The models used for object detection can be divided into two categories: Two-stage detection frameworks and One-stage (Unified) detection frameworks. In the first, the process is divided into two phases. First, there is a proposal for candidate regions of the image that may contain objects to be detected [6]. The classification is then made based on the first result, fine-tuning the regions, discarding false positives (for example, Faster R-CNN). One-stage detection frameworks perform the process at once, without the initial region proposal step and therefore allow a single model to be used, predicting the bounding boxes that contain the objects present, as well as the probabilities of these belonging to the classes considered [5] (for example the YOLO models). The YOLOv4 models analyze the image's features using different resolutions, maintaining the original image's height/width ratio. These models manage, by adapting the size of the initial anchors to the specific dataset, to detect objects of various scales in the images [6] and allow the correct detection of overlapping objects of different classes [3]. For this reason, they present advantages when used for this problem, since it is common to have the presence of smoke in the images where there is fire. However, these models have some disadvantages, such as the need to have a considerable amount of annotated images to obtain good results. As such, to solve the lack of annotated data, two datasets are proposed with the annotation of Fire and Smoke class objects for training and model evaluation. These datasets were used to optimize the results in the context of forest fires. In addition, the transfer learning technique was also used, with a pre-trained model with the Imagenet dataset. The initial weights resulting from the pre-training were kept, responsible for the extraction of more low-level features. The last layers of the model, responsible for the extraction of features specific to the problem, were retrained [6]. The use of transfer learning makes the training process less time-consuming and improves the model's learning capacity.

[1] Project PCIF / MPG / 0128/2017, FireLoc - Where's the Fire? - Identification, positioning, and monitoring forest fires with crowdsourced data

3 Proposed approach

For the development of the fire and smoke detection system, an object detection approach was adopted, using YOLOv4 [3] models. These models detect the specific location of fire and smoke in the images. Therefore, they require an indication of where the objects are present within the image, using bounding boxes. To perform the manual annotation of the training and testing datasets proposed, according to the YOLO annotation format, the tool LabelImg was used.

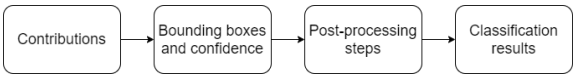


Figure 1: Detection system steps

Figure 1 shows the sequence of steps necessary to detect fire and smoke in the images submitted by the application users. The model first identifies the parts of the images that contain fire or smoke with bounding boxes, and the corresponding confidence score associated with each detected object. The classification results are then obtained with the post-processing step, which allows integration with the FireLoc system. For the classification results, the **Fire**, **Smoke**, and **Neutral** classes are considered. The images in which fire is detected belong to class **Fire**, and the images in which smoke is detected, and no fire is detected belong to class **Smoke**. The remaining images, in which no object is detected, belong to the Neutral class.

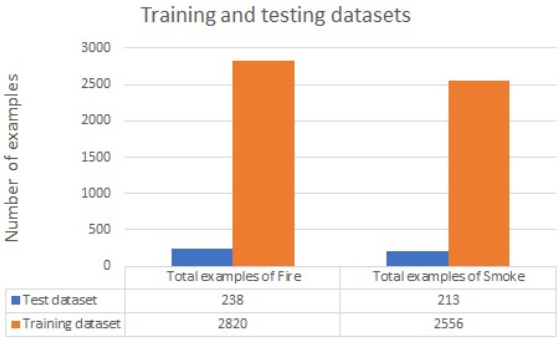


Figure 2: Number of object examples in each dataset

For the training of these models, an open-source dataset from [1] was used. This dataset contains an equal amount of images from each class: **Smoke**, **Fire**, and **Neutral**). It contains 1000 images of each category and is named *Fire-Smoke-Dataset*. To complement this training dataset

approximating their characteristics and test the models' performance in the context of forest fires, an image dataset taken in a simulacrum[2] performed by firefighters was also used. The pictures were taken using different smartphones and tablets and correspond to a context similar to that of using the FireLoc application. These datasets were annotated according to the YOLO annotation format, resulting in the distribution of examples shown in the graph of figure 2.

4 Results and Discussion

For the evaluation of YOLOv4 models' performance, the IoU threshold was set to 0.3, lower than the typical 0.5 value used int the COCO dataset detection competition. The lowering of the IoU threshold used was intended to increase the tolerance for object location errors, obtaining a greater number of valid detections. The results of object detection in the images obtained by the models were evaluated by calculating the mean average precision (mAP) value obtained and the classification results of the images as a whole, using confusion matrices. One of the reasons for the analysis of mAP is that, by considering the level of confidence in the detections made, it is possible to evaluate the relationship between false positives and false negatives [4]. The analysis of mAP results allows the evaluation of the models' detection performance. The datasets proposed for training and testing were used for training the models, using the default input size for the images, 416x416. Two classes of objects present in the images were considered: Fire and Smoke, and the anchors were adjusted to the training dataset, using k-means, to approximate the dimensions to those of the objects present in the test images.

Additional tests were performed to adjust the confidence threshold. The model with the best performance obtained in the previous test was used, varying the confidence threshold between 5%, 10%, 15%, 20%, and 25%. All detections whose confidence indicated by the model is lower than the value of the defined threshold are discarded and, as such, the adjustment of the defined threshold may have an influence on the results of both fire and smoke detection in the images (that is, the ability to detect the location in the space of the images correctly) as well as the classification of the images as a whole. Before this test, the confidence threshold value used was the same as in the YOLOv4 [3] work. The graph in table 1 shows that the lower the defined confidence threshold, the best mAP result.

Confidence threshold	**5%**	**10%**	**15%**	**20%**	**25%**
mAP results	**62.2**	60.3	**57.5**	54.8	52.22

Table 1: Test mAP results

These results indicate that the lower the confidence threshold, the greater the number of detections performed, resulting in better mAP results. Therefore, although the best result is obtained with 5% confidence, detection performance was compared considering 5% and an intermediate point 15% confidence.

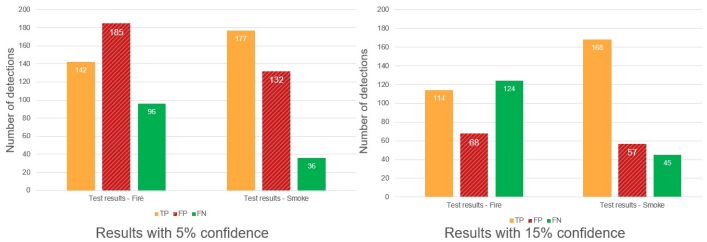


Figure 3: Test results with different confidence threshold

	Predictions – confidence 5 %		
Ground-truth	**Fire**	**Neutral**	**Smoke**
Fire	78	0	6
Neutral	1	22	14
Smoke	14	0	80

Table 2: Classification results with 5% confidence

The results presented in the confusion matrices in tables 2 and 3, correspond to the classification with three classes, after the application of the first post-processing step. The results obtained with 5% confidence,

[2]Images collected on May 15, 2019, in tests carried out in Serra da Lousã by *Associação para o Desenvolvimento e Aerodinâmica Industrial (ADAI)*

	Predictions – confidence 15 %		
Ground-truth	**Fire**	**Neutral**	**Smoke**
Fire	67	0	17
Neutral	0	31	6
Smoke	6	5	83

Table 3: Classification results with 15% confidence

in graph 2, show that the increase of false positive detections does not have a significantly negative impact on the classification results, with the model correctly identifying all the dataset's images where fire or smoke is present. However, as a consequence of the increase of the number of detections that are considered valid, the number of false positives also increases, reaching a total of 15 Neutral images in which smoke or fire is mistakenly identified.

Therefore, two points of operation for the system are proposed, alternating the threshold of confidence in the detections between 5% and 15%, depending on the number of submissions made by the application users. The possibility of switching between two operating points allows for a larger or smaller filter to be applied to the submitted images, adapting the system's operation to times of greater or lesser occurrence of forest fires.

Figure 4: Example of detection results

When comparing the detection results obtained in the same image considering the two proposed operating points, in figure 4, it is possible to observe repeated detections of the same object in both results. When using a 5% confidence threshold, the number of repetitions is noticeably higher, resulting in a greater number of false negatives in the detection results (in the graphs of figure 3). In this case, independently of the defined operation point, the image would be correctly classified with class **Fire**.

5 Conclusions and Future Work

In conclusion, the results obtained for the fire and smoke detection problem in static images are promising. With the developed system, it is possible to get the fire or smoke location detected in the image's space. The YOLO models also allow fire and smoke detections to be made on video, which can be useful in integrating with the FireLoc system if video submissions, in addition to still images, are accepted. The system can be dynamic, absorbing new information by applying new cycles of additional training, using images submitted by users, adapting it better to the application's operating context.

References

[1] Fire-Smoke-Dataset. `https://github.com/DeepQuestAI/Fire-Smoke-Dataset/releases/download/v1/FIRE-SMOKE-DATASET.zip`, note = accessed 2020-08-22, .

[2] FireLoc - Localize o Fogo. `https://fireloc.org/`, note = accessed 2020-09-06, .

[3] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. YOLOv4: Optimal Speed and Accuracy of Object Detection. (May), 2020.

[4] Mark Everingham, Luc Van Gool, Christopher K.I. Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (VOC) challenge. *International Journal of Computer Vision*, 88(2):303–338, 2010. ISSN 09205691.

[5] Licheng Jiao, Fan Zhang, Fang Liu, Shuyuan Yang, Lingling Li, Zhixi Feng, and Rong Qu. A survey of deep learning-based object detection. *IEEE Access*, 7(3):128837–128868, 2019. ISSN 21693536.

[6] Li Liu, Wanli Ouyang, Xiaogang Wang, Paul Fieguth, Jie Chen, Xinwang Liu, and Matti Pietikäinen. Deep Learning for Generic Object Detection: A Survey. *International Journal of Computer Vision*, 128(2):261–318, 2020. ISSN 15731405.

[7] Zhong Qiu Zhao, Peng Zheng, Shou Tao Xu, and Xindong Wu. Object Detection with Deep Learning: A Review. *IEEE Transactions on Neural Networks and Learning Systems*, 30(11):3212–3232, 2019. ISSN 21622388.

Breast MRI Multi- Sequence Segmentation and Registration

João F. Teixeira[1,2]
jpfteixeira.eng@gmail.com

Sílvia Bessa[2]
silvia.n.bessa@inesctec.pt

Hélder P. Oliveira[2,3]
helder.p.oliveira@inesctec.pt

[1]Faculdade de Engenharia
Universidade do Porto, Porto, Portugal

[2]INESC-TEC
Porto, Portugal

[3]Faculdade de Ciências
Universidade do Porto, Porto, Portugal

Abstract

Different sequences of the same medical exam, as for instance MRI's *T1w* and *Dyn*, display different image features that enable the segmentation of specific objects to be easier in one over the other. In breast cancer research, *T1w* addresses better the diverse breast anatomy, while *Dyn* outshines over lesion segmentation. The present study proposes a methodology to tackle an unapproached task, in order to facilitate the volumetric alignment of data retrieved from *T1w* and *Dyn* sequences, leveraging breast surface segmentation and subsequent registration. The process seems to have promising results as average two-dimensional contour distances are at sub-voxel resolution and visual results seem well within range for the valid transference of other segmented or annotated structures.

1 Introduction

Magnetic Resonance Imaging (*MRI*) is often performed on breast cancer patients and allows 3D image reconstruction of the breast as it captures regular interval slices from the patient's torso. Image processing methods are often used to analyse these challenging *MRI* sequences, namely focusing on T1-weighted (*T1w*) and MRI Dynamic Contrast-Enhanced (MRI-DCE, *Dyn* henceforth). Specific objects in the torso are enhanced on particular sequences and so, it stands to reason focusing the development efforts towards dealing with the clearer options first.

The present study is an adaptation of [5], which aims to automatically obtain the breast anterior surface on *T1w* and on *Dyn* at instant zero (*Sd0*). Also, to join lesion annotations with the remaining anatomy reference points, the segmented surfaces from *T1w* and *Sd0* are registered using a simplification of the Iterative Closest Point algorithm (ICP).

1.1 Related Work

Across multiple cancer specificities, and breast cancer research in particular, there is a significant amount of data and methods associated with segmentation and multi-modal registration or fusion.

Segmentation approaches range from Maximum a Posteriori Estimation approaches, Expectation Maximization–Markov Random Field techniques and Atlas-based approaches to U-Net methodologies. However, research does not focus as much on *T1w* or *T2w* sequences as it does on *Dyn*, due to the lower difficulty of lesion segmentation on that sequence. Furthermore, those segmentation procedures are largely directed only towards the lesion and generally customarily the breast surface.

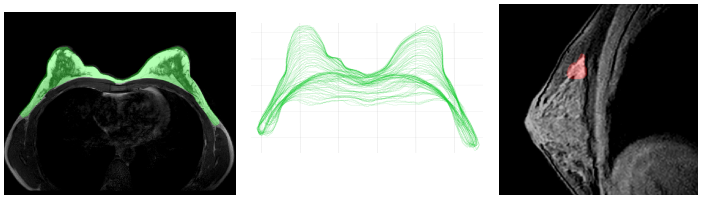
Breast multi-modal registration tasks generally involve fusing the *Dyn* sequences to other entirely different modalities, such as PET and CT [1], or between the 3D three-dimensional (3D) MRI and 2D data such as X-ray Mammography [3]. There are also human biology based works that focus on intra-modality alignment, commonly associated with the monitoring of some disease's progression, as usually found applied to brain CT scans [4]. Nevertheless, some similar work is also done with breast *Dyn* sequences [2].

However, this paper seems to be among the earlier work concerning *MRI* intra-patient registration between *T1w* and *Dyn* or derived subsequences, as in *Sd0*. Hence, we start studying low complexity approaches such as edge detection, contour refinement and rigid registration.

2 Dataset

The dataset used was provided by the *Redacted* project and consists of *T1-weighted* thoracic MRI exams (*T1w*) from 27 breast cancer patients, obtained with a Philips Ingenia 3.0T MRI scanner. Each exam comprises 60 gray-scale axial images, with the approximate dimensions of 3 mm thickness and resolution of 720×720 pixels (0.3-0.5 mm/pixel). Additionally, each *T1w* acquisition has a corresponding dynamic contrast study (*Dyn*) that includes the sequence data at the instant zero (*Sd0*). In turn, this sequence comprises 300 gray-scale sagittal images, with 1 mm thickness and a resolution of 300×300 pixels (0.5-0.6 mm/pixel), in a narrower field of view.

Due to the specific ease of annotating, the *T1w* also has available binary masks for the breast and the *Sd0* has the lesion we aim to transport (Figure 1). All annotations were manually performed by experts with more than 5 years of experience.



(a) *T1w* breast (b) *T1w* breast perimeter (c) *Sd0* Lesion

Figure 1: Annotation Images

3 Methodology

The tasks intended to be tackled include *T1w* and *Sd0* sequences' segmentation and subsequent registration of both segmented breast surfaces. For the segmentation task, *T1w* and *Sd0* volumes are individually processed using the same pipeline, despite their difference in field of view, voxel resolution and extent. *Sd0* volumes are rotated in 90 degrees, so that both anatomical volumes face the same direction, on the axial view.

3.1 Breast Surface Segmentation

The approach tries to enclose the breast surface in a solid region, ignoring internal structures when possible, enabling a smooth generation of the breast surface. The segmentation pipeline is shown in Figure 2.

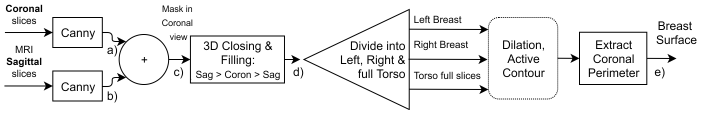
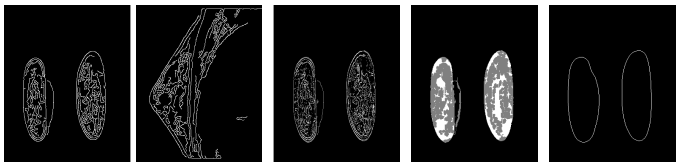


Figure 2: Segmentation Pipeline for one MRI sequence

The initial step is a Canny detector for obtaining the salient volume boundaries. This is performed in both coronal and sagittal directions (Figures 3a, 3b), to compensate potential gaps in perspective, namely hidden information of the inframammary folds. The edge maps are joined (Figure 3c) and closed with a 3D sphere. Flood-filling operations ensue: along the sagittal perspective, then along the coronal and again along the sagittal view (Figure 3d). Next we applied dilation filtering and a Chan-Vese level-set block, over the coronal view. This processing is done individually on separated left and right breast images, until a single object is found on the filling step. This avoids fusing the breasts on the active contour step. Lastly, the surface perimeter is then extracted (Figure 3e). An example of the 3D segmentation extents for a full patient is shown in Figure 4.

3.2 Registration

First we convert the coordinates of the point clouds to real world values (voxel resolution is applied to the point list). A point cloud is rotated



(a) Canny Coronal (b) Canny Sagittal (c) Mask Coronal (d) Filling (e) 3D Coronal Perimeter

Figure 3: Pipeline intermediate results for *Sd0*

so both (*Sd0* and *T1w*) face the same orientation. The point clouds are subsequently processed by an ICP algorithm, that imposes a no rotation restriction, as both sequences of data are captured during the same session and with the patient lying down, trying not to move. This will enable to accurately convert the desired spatial data - in this case the lesion's points acquired on the *Sd0* - and place it on the respective location, among the structure data on the base view (here, the *T1w*). An example of the process is shown in Figure 5.

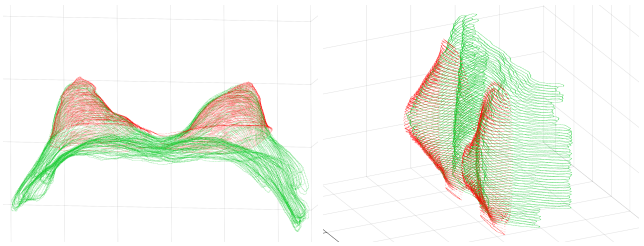
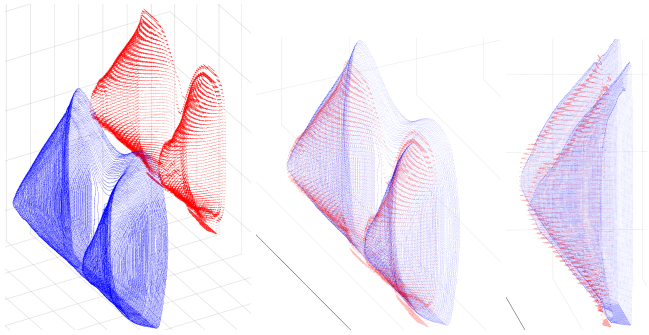


Figure 4: *T1w* Segmentation output (red) against GT contour (green)



(a) Pre-Registred (b) Registred (c) Registred

Figure 5: Registration step. *T1w* (red), Registered *Sd0* (blue)

3.3 Evaluation metrics

For segmentation we can only fairly compare the *T1w* segmentation with the *T1wGT*, as there is no breast GT mask for *Sd0*. Furthermore, it should only be done in one direction, as *T1wGT*'s perimeter also includes the posterior breast contour (Fig 1b). Registration wise, the metric is obtained only from the aligned surfaces as using the *T1wGT* contour could negatively impact the registration process due to its extensive perimeter.

In neither case, area metrics can be provided as the segmented frontal breast surfaces are not comparable to the solid *T1wGT* objects. Hence, the metrics performed are *Average Distance* (one way) and *AD* (bidirectional). Extreme valued cases across all patients are also shown.

4 Results and Discussion

First, we observed an increasing surface distance as we reached the vertical extremities (Figure 4). It was expected as less homogeneous intensities are found in the top and bottom slices. In turn, this influences the segmentation method which operates on the sagittal and coronal perspectives and slightly expands the real boundaries to obtain a smoother, closed result.

The registration process of the segmented surfaces (Figure 5) presents acceptable alignments, where larger mismatches may be attributable to segmentation errors at the vertical extremities, in particular of the *T1w* surface. The breast shape acquired from both sequences seems to match enough to conduct reliable registration, producing an average error around the central slices of approximately the 4mm.

When comparing to *T1wGT*, the *Sd0* follows closely the outer skin interface, while the *T1w* segmentation follows the inner one, in many cases having the manual annotation averaging between both.

Table 1 presents numerical evidence of the results. The 3D *T1w* segmentation results point to about a 4 to 7 pixel error, which is a good result, considering some of the method's limitations and the reliability of the dataset.

Table 1: Segmentation and registration 3D errors

	Min	Avg. Dist.	AD	Max
T1w to *T1wGT*	1.23	2.20 (0.47)	*n.a.*	3.05
Sd0 to *T1w*	1.75	2.91(1.30)	2.57 (1.01)	6.17
T1w to *Sd0*	1.37	1.77 (0.47)		3.76

All metrics in mm (min is best). Avg. Dist. and AD present values averaged across all patients, and respective standard deviations in parenthesis.

On the registration side, a trend of larger *Sd0* to *T1w* error was anticipated and verified, as the *Sd0* has roughly 3 times more slices than *T1w*, for the same vertical extent. General errors are quite low considering that *T1w* has a slice thickness of 3 mm. The Avg. Dist. values are close, confirming that both segmentation surfaces have fairly similar shapes and extents. This validates the balanced method performance across both sequences. This also argues for the capacity of this ICP setup for the intended objective. Finally, and naturally, AD middles the Average Distance of each sequence, leaning more on the *Sd0* to *T1w* direction, as *Sd0* tends to have more points on the clouds than *T1w*.

5 Conclusion

An early pipeline for the untried fusion between the breast outer contour of *T1w* and *Sd0 MRI* sequences has been proposed. The main objective was to unify both sequences under the same orientation and 3D space, to combine both sources of annotations.

For this, a breast surface segmentation plus registration approach was employed. Both visual and quantitative outcomes show encouraging results, managing average contour distances below *T1w*'s slice thickness value. Although the approach may require further adjustments for an atlas development and other goals, the proposed pipeline seems to fulfill the purpose of joining the annotations of these two sequences.

Acknowledgements: This work is financed by National Funds through the Portuguese funding agency, FCT - Fundação para a Ciência e a Tecnologia within project UIDB/50014/2020, and PhD grants number SFRH/BD/135834/2018 and SFRH/BD/115616/2016.

References

[1] I.D. Dmitriev, C.E. Loo, W.V. Vogel, K.E. Pengel, and K.G.A. Gilhuijs. Fully automated deformable registration of breast DCE-MRI and PET/CT. *Phys. Med. Biol.*, 58(4):1221–1233, 2013.

[2] Y.C. Gong and M. Brady. Texture-Based Simultaneous Registration and Segmentation of Breast DCE-MRI. In *Digital Mammography*, pages 174–180, 2008.

[3] T. Hopp, P. Baltzer, M. Dietzel, W.A. Kaiser, and N.V. Ruiter. 2D/3D image fusion of X-ray mammograms with breast MRI: visualizing dynamic contrast enhancement in mammograms. *J Comput Assist Radiol Surg*, 7(3):339–348, 2012.

[4] A. Klein, J. Andersson, B.A. Ardekani, J. Ashburner, B. Avants, M. Chiang, G.E. Christensen, D.L. Collins, J. Gee, P. Hellier, J.H. Song, M. Jenkinson, C. Lepage, D. Rueckert, P. Thompson, T. Vercauteren, R.P. Woods, J.J. Mann, and R.V. Parsey. Evaluation of 14 nonlinear deformation algorithms applied to human brain MRI registration. *Neuroimage*, 46(3):786 – 802, 2009.

[5] J.F. Teixeira, S. Bessa, P.F. Gouveia, and H.P. Oliveira. A Framework for Fusion of T1-Weighted and Dynamic MRI Sequences. In *Pattern Recognit. Image Anal.*, pages 157–169, 2020.

Sarcopenia Diagnosis: Deep Transfer Learning versus Traditional Machine Learning

Carlos Sobral
carlosfssobral@gmail.com

Jose Silvestre Silva
jsilva@ci.uc.pt

Alexandra Andre
alexandra.andre@estescoimbra.pt

Jaime B. Santos
jaime@deec.uc.pt

University of Coimbra, Department of Electrical and Computer Engineering, Portugal

Portuguese Military Academy and Military Academy Research Center (CINAMIL), Lisbon, Portugal
Laboratory for Instrumentation, Biomedical Engineering and Radiation Physics (LIBPhys-UC), Coimbra, Portugal

Coimbra Health School, Coimbra, Portugal

University of Coimbra, CEMMPRE, Department of Electrical and Computer Engineering, Portugal

Abstract

Sarcopenia is a syndrome characterized by progressive and generalized loss of skeletal muscle mass and muscle strength. Through the analysis of ultrasound images, this work compares the effectiveness between traditional deep transfer learning and three traditional classifiers, FineKNN, CubicSVM and SubspaceKNN. The results showed that the deep learning transfer had the best final classification, 98.3%; the traditional classifier that presented the better performance was CubicSVM, with an efficiency of 97.9%, higher than the others, with SubspaceKNN achieving 97.1% and FineKNN reaching 96.5%.

Keywords: Sarcopenia, ultrasound, traditional classifiers, Deep Transfer Learning, Inception-v3, FineKNN, CubicSVM, KNN Subspace.

1 Introduction

Deep learning is a method that has the ability to "learn" automatically the mid and high-level features from untrained images. Since the creation of AlexNet, the winner of the "ImageNet" Large Scale Recognition Challenge" (ILSVRC) in 2012 [1], deep learning attracts attentions in the field of machine learning [2]. In 2013, deep learning was selected for the top 10 of the most innovative technologies [3]. Nowadays, this technology is successfully applied in several areas, such as medical imaging [4], using different anatomical structures, such as muscles [5].

Musculoskeletal tissue has as main task the contraction and energy production to carry out movement. It is composed by elongated cells, and at the level of its histology has a striated appearance. These cells have transverse striations, alternated by a light band and a dark band, containing inside the main components for the process of contraction and muscle relaxation, called sarcomeres. As time goes by, sarcomeres are degraded leading to sarcopenia.

In 2010, the European Working Group on Sarcopenia in Older People (EWGSOP) defined sarcopenia as "a syndrome characterized by progressive and generalized loss of skeletal muscle mass and muscle strength, with risk of adverse effects such as physical disability, poor quality of life and death" [6].

2 Methodology

This section aims not only to make a brief explanation of each step performed throughout the project, but also to indicate the workflow of it. The flow and the steps performed are illustrated in the following figure.

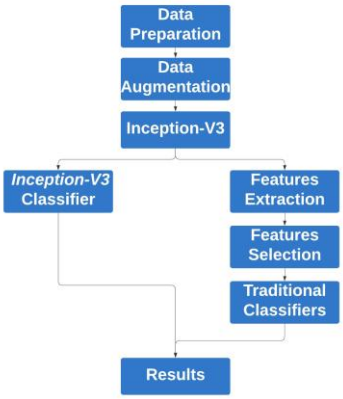


Figure 1: WorkFlow.

2.1 Data Preparation

From acquired images, several 40x40 ROI were extracted, since ROIs of larger size captured undesirable elements such as aponevroses. Once a rotated rectangular ROI presented the corners in black, which could negatively influence the final results, it was used circular ROIs to perform the rotations in the data augmentation process. In order to use this ROIs in the neural network, a resize was performed, changing its dimension to 299x299x3, since these are the input dimensions required by inception-v3 network. The ROI was replicated three times to fill the three layers of the RGB image.

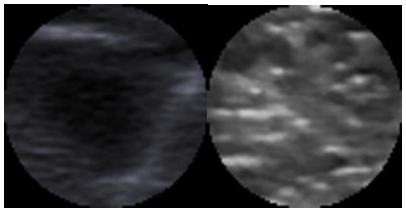


Figure 2: (a) Normal muscle ROI (b) Sarcopenic muscle ROI.

2.2 Data Augmentation

This is a mandatory process to be carried out when talking about Deep Transfer Learning and Machine Learning, since both need large amounts of data to be able to produce good accuracy. For the tests performed, three mathematical operations were used.

a) Rotations: 3 different angles were defined, being 10°, 20° and 30°. For this purpose, the matlab command "imrotate" was used. Having in mind these values, each image generated six new images.

b) Noise: In this operation was used the Gaussian noise formula in matlab:

$$x = b + \sqrt{c} * rand(size(b)) \quad (1)$$

In this expression b is the image (ROI) and c is a chosen value from one of three distinct values that were randomly selected. In this case 6,12 and 25 were used. Then, three new images were created.

c) Rotations + Noise: 18 new images were created, all obtained from the merging of rotation operations (clockwise and non-clockwise) and the addition of noise.

d) Flip: once the percentages of the training, validation and testing data were chosen, the number of images was doubled in each one steps with the "flip" operation, mirroring the images. For this, the matlab command "fliplr"was used.

2.3 Inception-V3

The neuronal network Inception-V3 was used for the analysis of the images. It has an architecture based on 2 types of factorization.

2.3.1 Factorization into small convolutions

The factorization of convolutions aims to reduce the number of connections/parameters without affecting the accuracy of the network. This allows the reduction of the size from convolution layers, namely a 5x5 convolution, which is replaced by two 3x3 convolutions. That leads to parameters reduction [7].

2.3.2 Factorization into Asymmetric Convolutions

This factorization uses the layers resulting from the previous process. These layers have their size reduced, namely a 3x3 layer, which is converted into two 1x3 and 3x1 layers. Using a 3×1 convolution followed by a 1×3 convolution is the equivalent of sliding a two-layer network with the same filter size as in a 3×3 convolution.

In order to reduce the computational costs of the network, it was found that the transformation into 1x7 and 7x1 layers provided better results [7].

2.3.3 Convolutional Layers

A convolutional layer contains a set of filters whose parameters need to be learned. Each filter is concatenated with the input volume to calculate an activation map, i.e., the filter goes through the entire width and height of the input. Finally, the product of the points between the input and the filter are calculated at each spatial position.

2.3.4 Pool Layers

The Pool layer is added after the convolutional layer. Specifically, after a nonlinearity (for example, ReLU) has been applied to the output of features maps by a convolutional layer. There are 2 types of pool layers:

a) Average pool: Calculates the average value of each patch on the features map.

b) Max pool: Calculates the maximum value for each patch of the features map.

The result of using this type of layers and creating sampled feature maps is a summarized version of the features found in the input.

2.5 Features Extraction

Loading dataset into the inception-V3 network, the "avg_pool" layer of this network was used to obtain the 2048 features of each image. Using the “activations” command present in Matlab it was possible to create an array with all the extracted features. Then this array was introduced into the feature extractor toolbox *FEAST* in order to select the best 400 features. This toolbox has several algorithms to perform the detection of the best features, to be used in the training of traditional classifiers.

3 Results

Datasets were obtained from 144 images (ROIs of the original images, 61 normal muscles and 83 sarcopenic muscles). The percentages assigned to the training, validation and testing steps were the following:

a) dataset A: 60% for training, 20% for validation and 20% for testing, which corresponds to 7200 images (4300 training, 1450 validation and 1450 test).

b) dataset B: 70% for training, 15% for validation and 15% for testing, which corresponds to 7200 images (5000 training, 1100 validation and 1100 test).

c) Dataset C: 80% for training, 10% for validation and 10% for testing, which corresponds to 7200 images (5750 training, 750 validation and 700 test).

Finally, three new datasets were obtained from 285 images (2 ROIs per original image), but 3 of the images only allowed to select 1 ROI since they have small muscles . The percentages assigned to the training, validation and testing steps were the following:

a) A2R dataset: 60% for training, 20% for validation and 20% for testing, which corresponds to 14350 images (8550 training, 2900 validation and 2900 test).

b) B2R dataset: 70% for training, 15% for validation and 15% for testing, which corresponds to 14350 images (9950 training, 2150 validation and 2150 test).

c) C2R dataset: 80% for training, 10% for validation and 10% for testing, which corresponds to 14350 images (11400 training, 1450 validation and 1450 test).

We used three different proportions of training, validation and test in order to verify the effect of the training percentages on the final results.

Tests were carried out on several traditional classifiers presented in the “classification learner” application of Matlab to identify which classifier would produce the best performance.

As input, these classifiers received the matrix with the features extracted from all the training images in the dataset and a matrix with the respective labels.

After training the classifiers, the results achieved by the best three classifiers were exported to be used later, in order to evaluate the test images of the dataset in question.

For the analysis with the inception-V3 network, the following values were chosen for the different hyper-parameters: Epochs = 5; LearningRateFactor=0.001;MiniBatchSize=100; WeightLearnRateFactor = 10; BiasLearnRateFactor = 5.

Method	Dataset A	Dataset B	Dataset C	Dataset A2R	Dataset B2R	Dataset C2R
CubicSVM	95.0%	93.3%	90.7%	97.1%	97.9%	97.9%
Fine KNN	91.9%	91.5%	89.6%	95.4%	96.5%	96.5%
SubspaceKNN	92.9%	92.1%	90.3%	95.9%	97.1%	96.4%
Inception-V3	98.3%	93.3%	90.9%	97.9%	98.0%	97.9%

Table 1: Accuracy values achieved by the four implemented methods.

Looking closely to the obtained results, it was concluded that traditional classifiers performance is lower than deep transfer learning since in datasets with only one ROI per image, the maximum classification value achieved by traditional classifiers was 95%, i.e., a difference of 3.2% compared to the best result obtained by inception-V3. With regard to datasets with double ROI per image, the classifications between both methods are similar. However, only Inception-V3 achieved an efficiency of 98%. These results suggests that deep transfer learning has a superior performance when compared to traditional classifiers.

3 Conclusions and Future Work

The datasets of this project were obtained through 144 original images and since both deep transfer learning and machine learning needs a large amounts of data, the acquisition of more images related to patients with and without the pathology can also allow an increase on accuracy. One of the aspects that can be studied in future work is the size defined for the ROIs, since having these with smaller sizes can also produce better results.

Acknowledgments

This research is sponsored by FEDER funds through the program COMPETE – Programa Operacional Factores de Competitividade – and by national funds through FCT – Fundação para a Ciência e a Tecnologia –, under the project UIDB/00285/2020.

References

[1] Teofilo F. Gonzalez. Handbook of approximation algorithms and metaheuristics. Handb. Approx. Algorithms Metaheuristics, pages 1–1432, 2007.

[2] Shengfeng Liu, Yi Wang, Xin Yang, Baiying Lei, Li Liu, Shawn Xiang Li, Dong Ni, and Tianfu Wang. Deep Learning in Medical Ultrasound Analysis: A Review. Engineering, 5(2):261–275, 2019.

[3] Ge Wang. A perspective on deep imaging. IEEE Access, 4:8914–8924, 2016.

[4] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Fran-cesco Ciompi, Mohsen Ghafoorian, Jeroen A.W.M. van der Laak, Bram van Ginneken, andClara I. Sánchez. A survey on deep learning in medical image analysis. Med. Image Anal., 42:60–88, 2017.

[5] P. Burlina, N. Joshi, S. Billings, I. J. Wang, and J. Albayda. Unsupervised deep novelty detection: Application to muscle ultrasound and myositis screening. Proc. - Int. Symp. Biomed. Imaging, 1910–1914, 2019.

[6] Alfonso J, et al. Sarcopenia: Europeanconsensus on definition and diagnosis. Age Ageing, 39(4):412–423, 2010.

[7] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethin-king the Inception Architecture for Computer Vision. Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., 2818–2826, 2016.

Multispectral Images Applied to Face Recognition

Luis Lopes Chambino
luis.chambino@tecnico.ulisboa.pt
Military Academy, Portugal,
Instituto Superior Técnico, Universidade de Lisboa, Portugal

José Silvestre Silva
jose.silva@academiamilitar.pt
Military Academy & CINAMIL, Lisbon, Portugal,
LIBPhys-UC, Coimbra, Portugal

Alexandre Bernardino
alex@isr.tecnico.ulisboa.pt
Institute for Systems and Robotics (ISR), Portugal
Instituto Superior Técnico, Universidade de Lisboa, Portugal

Abstract

Facial recognition is a method of identifying or authenticating the identity of individuals through their faces. Systems that use multispectral images in face recognition obtain better results that those who use only visible images. In this work, we propose a multi-channel deep convolutional neural network approach for facial recognition using multispectral images. A study is carried out to assess the performance of Support Vector Machines and k-Nearest Neighbor classifiers to classify the 256-d embeddings obtained by adapting the Domain Specifc Units in the LightCNN. Experimental results in the Tufts face dataset show competitive performance in facial recognition obtaining a rank-1 score of 99.5%.

1 Introduction

Nowadays it is possible to see a growth of applications that use facial recognition systems, whether for collective use, as in companies, or for personal use, as in smartphones. There is also an increasing usage of more than one spectral range to improve results in facial recognition.

There are two main modes of image acquisition in facial recognition systems: in a controlled environment, where a person cooperates in acquiring images, and in an uncontrolled environment, also known as "in the wild", where a person does not cooperate or has no knowledge during the phase of image acquisition. Systems that use only the visible spectrum (VIS) have several obstacles, such as occlusions, pose variation, non-cooperation of the person and, the most problematic, changes in the luminosity. As a result, it is necessary to complement these facial recognition systems, either with the use other biometric sensors (e.g. fingerprint or iris) or other spectral bands, in order to minimize these problems.

The infrared spectrum, namely the Near Infrared (NIR), Short Wave Infrared (SWIR), Medium Wavelength Infrared (MWIR) and Long Wavelength Infrared (LWIR) spectral bands, has been used successfully in facial recognition systems, as a complement of the visible spectrum [1]. These systems, which use more than one spectral band, are called multispectral.

The infrared spectral band has several advantages when compared to the visible spectrum; it is imperceptible to the human eye and, at the same time, less sensitive to differences in luminosity. For instance, the night cameras used in video surveillance have LEDs with emission in the infrared spectrum to illuminate the scene and perform night surveillance without people realizing it.

The spectral bands NIR and SWIR are very close to the visible spectrum, thus afford an easy adaptation of automatic learning methods trained with images of the visible spectrum. The MWIR and LWIR spectral bands (also known as thermal bands) allow the use of facial recognition systems at night, when the luminosity is very low or even zero.

Multispectral facial recognition systems, in comparison with only visible facial recognition systems, can be used as a method to add an extra security layer, to recognize a person more accurately, in accessing a high security place, in order to guarantee access only to authorized people. These places can be hospitals, schools, laboratories and military buildings [1].

Through the development of an improved facial recognition system, it is possible to guarantee a more reliable and more robust access control, protecting property and increasing people's safety.

2 Methods

Each multispectral image may have several channels, one for each spectral band. When an monospectral band is in RGB, this image is converted to greyscale, as it is a requirement in the next phases.

Face detection is performed in all channels using the OpenCV [2] deep neural network (DNN) to obtain a face bounding box. If face detection was inconclusive (normal in LWIR channels, since the DNN used was trained in VIS images) the face bounding box from the VIS channel is used.

After the face bounding box is obtained, face landmark detection is performed using the DNN provided by Dlib [3]. A face alignment is accomplished by transforming the image such that the eyes centres are horizontal and in a predefined coordinate.

Finally, the aligned images are resized to a size of 144 x 144 pixels, to perform data augmentation on the images in order to better generalize our model, later explained in section 2.3. Figure 1 shows images in each spectral band.

Figure 1: Illustrative images of the VIS, NIR and LWIR spectral bands.

2.1 Model Architecture

The method used to extract the face embeddings to perform facial recognition is based on the concept of Pereira et al. [4] (the Domain Specific Units (DSU)), and we use the LightCNN [5] as the deep convolutional neural network, in contrary with Pereira work where he uses the Inception-ResNet-V2 [6] neural network.

Pereira et al. [4] showed that low level features in deep convolutional neural networks can be adapted to satisfy a specific spectral band, doing so it is not necessary to re-train the entire convolutional neural network. For this reason, we use a neural network that was previously trained for the task of facial recognition.

The LigthCNN model used in this work was trained with several face images, in the visible band, for face recognition. This architecture was chosen because of the small number of parameters, when comparing with other networks in face recognition. The reduced set of parameters was achievable because of the use of the Max-Feature Map (MFM) activation function as an alternative for the Rectified Linear Units (ReLu), which supresses low activation neurons in each layer [5].

The proposed architecture is shown in the Figure 2. The LightCNN takes as input a 128 x 128 pixels image and produces a 256-dimensional embedding, which can be used as a face representation. Each 256-d embedding represent the identity of the person through the spectral band of the channel used. Then all the embeddings produced by the channels are concatenated. In the last fully connected layer, the linear activation function was used.

The last fully connected layer it is added to produce the final 256-dimensional embedding, which can be used as a face representation by all the channels.

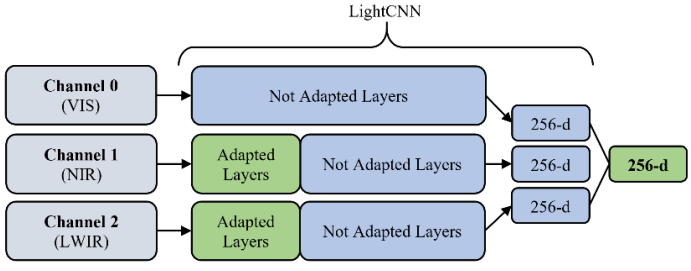


Figure 2: Overview of the proposed architecture: adapted layers (green) and not adapted layers (blue).

Using the pre-trained weights from the LightCNN, trained for face recognition in the visible band, it is possible to avoid a possible overfit, since the available multispectral datasets have a very limited number of images.

Obtained the 256-d embeddings it is now necessary to train a classifier to identify the person in the image. For comparison purposes, two classifiers were also used: the Support Vector Machines (SVM) and the k-Nearest Neighbour (kNN).

2.2 Dataset

The Tufts face dataset [7] was used to test the architecture. First it was necessary to clean and pre-process the dataset before using it. Initially the missing and corrupted images were excluded, then a facial detection was done. If facial detection was inconclusive a manual face detection was performed. Then all images where cropped and resized to a predefined size of 144x144 pixels. After this pre-processing task, the final dataset had 7 715 images from 109 persons.

This dataset was split into three subsets: 60% for training, 20% for validation and the last 20% for testing. It was performed a stratified split in the dataset so that each person is equally represented in each split. This step is necessary since the number of images per person in the dataset is not equal.

2.3 Training Procedure

Data augmentation was used to obtain a more generalized model. In the training set it was used random horizontal mirroring and random cropping to the size of 128x128 pixels. With the validation set it was only applied a center cropping to a size of 128x128 pixels, to comply with the required size of the LightCNN.

During the training procedure the proposed architecture was trained with the cross-entropy loss function. As the architecture was implemented in Pytorch, the cross-entropy loss function combines the Logarithmic SoftMax (LogSoftMax) and the negative log likelihood (NLLLoss) into a single function. Was used the Adam optimizer with a batch size of 16 and a learning rate of 10^{-3}.

3 Experiments and Results

To evaluate the performance of the proposed architecture several analyses were performed.

The first analysis allowed us to choose the appropriate classifier. Three classifiers were implemented: SVM with radial base function (rbf) kernel and the linear kernel and the kNN with the Euclidian distance. To obtain the best hyperparameters a stratified 5-fold cross-validation (CV) was performed during the training of each classifier with the training and validation set.

The hyperparameters fine-tuned were the regularization parameter (C), the kernel coefficient (γ) and the number of neighbours (k). The best hyperparameters, the range used in each hyperparameter, and the rank-1 value obtained with it are displayed in Table 1.

Table 1: Best hyperparameters obtained for each classifier.

Classifier	Hyperparameters			Rank-1
	$10^{-10} \leq C \leq 10^{5}$	$10^{-10} \leq \gamma \leq 10^{2}$	$1 \leq k \leq 25$	
SVM - Linear	> 0.01	-	-	99.8 %
SVM - rbf	10	10^{-4}	-	99.8 %
kNN	-	-	1	99.5 %

Determined the more suitable hyperparameters for each classifier they were trained only with the training set. Afterwards the classifiers were used to classify the 256-d embeddings from the test set. It was achieved a rank-1 score 99.70 %, 99.24 % and 99.24% for the SVM-Linear, SVM-rbf and kNN, respectively.

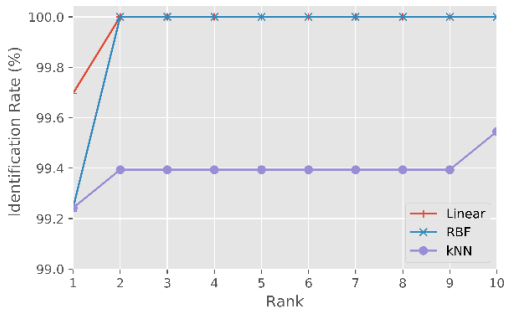


Figure 3: Cumulative matching curves for each classifier.

Figure 3 presents the cumulative matching curve (CMC) for each classifier up to rank-10. It is possible to observe that both SVM classifiers obtain a 100% score at rank-2 and have similar results. Further studies concluded that kNN obtains a 100% score at rank-102. These experimental results indicate that the SVM classifiers are more useful in identifying a person identity.

A comparison is made with other state of the art methods for face recognition that used the same dataset. This comparison can be seen in Table 2. When compared with other methods the proposed architecture proves to be a viable choice for multispectral face recognition, obtaining higher results.

Table 2: Comparison with state-of-the-art face recognition methods for the Tufts face dataset.

Method	Rank-1
Circular HOG [8]	94.5 %
TR-GAN [9]	88.7 %
Proposed methodology	99.7 %

4 Conclusions

Multispectral facial recognition still has plenty of space to evolve and improve. The main targets of multispectral facial recognition systems continue to be security and surveillance, especially in critical locations, such as airports or military classified areas.

In this work, it is proposed a new architecture for facial recognition using multispectral images. The architecture produces 256-d embeddings that represent the identity of a person through multispectral images. To test and compare this architecture it is used the Tufts face dataset. To classify the 256-d embeddings an SVM-linear classifier proved to be the best classifier, obtaining the higher rank-1 score. Experimental results verify the effectiveness of the proposed architecture in multispectral face recognition when comparing with other state-of-the-art methods.

Acknowledgements

This work was supported in part by the Military Academy Research Center (CINAMIL) under project Multi-Spectral Facial Recognition, and by FCT with the LARSyS – FCT Project UIDB/50009/2020.

References

[1] W. Zhang, X. Zhao, J. Morvan, and L. Chen, "Improving Shadow Suppression for Illumination Robust Face Recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, pp. 611–624, 2019.

[2] G. Bradski, "*The OpenCV Library*", Dr. Dobb's Journal of Software Tools, 2000.

[3] D. King, "*Dlib-ml: A Machine Learning Research*", Journal of Machine Learning Research, vol. 10, pp. 1755-1758, 2009.

[4] T. D. Pereira, A. Anjos, and S. Marcel, "Heterogeneous Face Recognition Using Domain Specific Units," *IEEE Transactions on Information Forensics and Security*, vol. 14, no. 7, pp. 1803-1816, Jul 2019.

[5] X. Wu, R. He, Z. Sun, and T. Tan, "A Light CNN for Deep Face Representation with Noisy Labels," *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 11, pp. 2884-2896, 2018.

[6] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-V4, Inception-ResNet and the Impact of Residual Connections on Learning," in Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence. San Francisco, United States of America: AAAI Press, 2017, p. 4278–4284.

[7] K. Panetta, et al. "A Comprehensive Database for Benchmarking Imaging Systems," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 3, pp. 509-520, 2018.

[8] S. Rajeev, K. Shreyas, Q. Wan, K. Panetta and S. Agaian, "Illumination Invariant NIR Face Recognition Using Directional Visibility ", *Electronic Imaging, Image Processing: Algorithms and Systems XVII*, 2019, pp. 273-1-273-7.

[9] L. Kezebou, V. Oludare, K. Panetta, and S. Agaian, "TR-GAN: thermal to RGB face synthesis with generative adversarial network for cross-modal face recognition", Proceedings SPIE, *Mobile Multimedia/Image Processing, Security, and Applications*, vol. 11399, 2020.

Artificial Intelligence in the Operating Room: evaluating traditional classifiers to predict patient readmission

Rita Sacramento
rita4sacramento@gmail.com
Instituto Superior de Engenharia de Coimbra

Rui Silva
rsilva@b-simple.pt
BSimple, Porto, Portugal

Inês Domingues
inesdomingues@gmail.com
Medical Physics, Radiobiology and Radiation Protection Group, IPO Porto Research Centre (CI-IPOP)

Abstract

The readmission of patients who had surgery is very prevalent. The goal of this work is to develop machine learning models to predict the likelihood of this readmission. The models will be based on several characteristics such as pathology, surgical speciality, surgical intervention, among others. Given a group of clinical cases, collected from 3 hospitals, the above mentioned parameters are collected and used to train and test machine learning algorithms such as Logistic Regression (LR), Support Vector Machines (SVM), K-Nearest Neighbour (kNN) and Decision Trees (DT). Data imputation and data balance techniques were also used. Models were developed with pre-surgery data only and also with data from after the surgery. Decision Trees have shown the best performance, having an accuracy of 91% before surgical intervention and an accuracy of 82% after surgical intervention.

1 Introduction

The analysis of the number of readmissions is of uttermost importance since, in addition to the added expenses for the hospital and the implications for the patient, it is also a marker of quality of the service provided by the healthcare facility. In a study carried out in Portugal between 2000 and 2008, it was possible to conclude that of the 5 514 331 unplanned hospitalisations, 4.1 % corresponded to hospital readmissions and that in episodes of readmission, hospital mortality was higher than in the remaining episodes, with the mortality rate in readmission episodes being 9.5 % and in the remaining 5.6 % [8].

The developed work consists in the application of machine learning techniques to a database of 21 112 occurrences, acquired in three different hospitals. Models were developed for two phases, before surgery, using a total of 7 attributes, and after surgery, using 13 attributes. Data imputation and data balance were performed and 4 classifiers were tested, Logistic Regression (LR), Support Vector Machines (SVM), K-Nearest Neighbour (kNN) and Decision Trees (DT). Decision Trees classifier is the one that has the best performance in both phases, achieving an accuracy of 0.91 before surgery and 0.82 after surgery.

2 Commercially available software

Companies like Jvion, AI Brisbane and Safecare AI use artificial intelligence to develop hospital software for decreasing or predicting readmission of patients.

Jvion [1] identifies and predicts patients at risk and defines actions to be taken for each patient. To reach a decision on the likelihood that a patient will be readmitted within 30 days, the data that is taken into account is not only clinical data, but also external data such as whether they have access to food, pharmacy or car. According to Jvion, self learning Eigen Spheres are used, although the details on this technique are not very clear [2]. According to the website, in a recent test, Jvion software was able to correctly identify patients at high risk of readmission 96% of the time.

AI Brisbane [3], focuses on forecasting using machine learning algorithms. It makes a selection and division of patients into groups, of high and low risk of readmission. In addition to the readmission forecast, the reason for this classification is also explained.

SafeCare AI [4] has a more preventive action, using real-time decisions where the focus is on actions that may decrease the likelihood of a next readmission. Medical data is processed using AI software to provide clinical decision support by intelligence emulation using machine learning, deep learning and artificial neural networks.

Unlike the above mentioned software, the software to be developed by BSimple is focused on the episode of the operation, assessing the risk of a certain patient being readmitted, before and after medical intervention.

3 Methods

This section will detail the steps taken during the development. The pipeline includes four main phases: pre-processing, training, evaluation and forecasting, as shown in Figure 1.

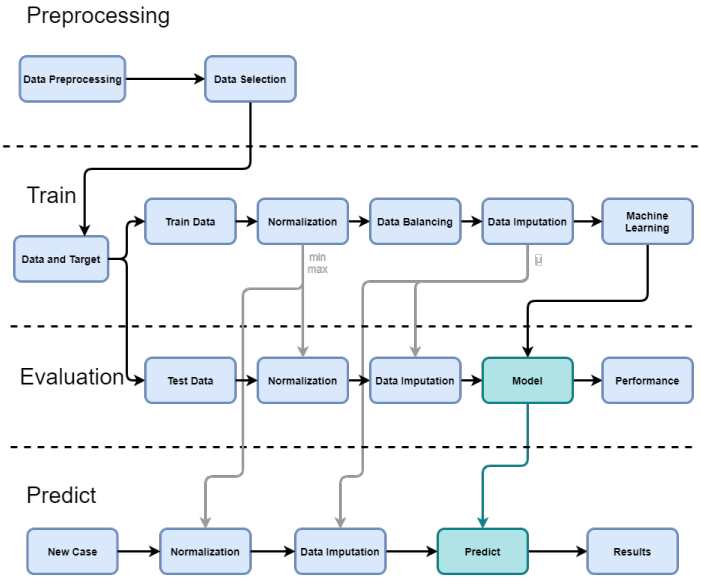


Figure 1: Methodology flowchart

Dataset: The used database has records from three Portuguese hospitals and contains 21,112 records. This database has 46 tables with different attributes. The attributes used in the model were empirically selected and depend if the instance corresponds to pre or post surgery.

Pre-processing: The database was first cleaned to remove all instances for which the surgery had been cancelled. To calculate which patients were readmitted, an interval of 30 days was considered.

Data selection: For the experiments before surgery, all cases with the variables: pathology, diagnoses, surgical speciality and surgical intervention with values equal to the case to be predicted were selected. These subgroups were then concatenated, removing the lines of repeated values. For the experiments after surgery, all cases with the variables: pathology, diagnoses, surgical speciality, surgical intervention, anaesthesia technique, and complications with values equal to the case to be predicted were selected. These subgroups were then concatenated, removing the lines of repeated values. This resulted in a set of 133 cases before surgery (79 non-readmissions and 54 readmissions) and 8 095 for after surgery (5 314 non-readmissions and 2 781 readmissions).

Data normalisation: The purpose of normalisation is to change the values of the data group used so that they all follow the same scale. The formula used is as follows: $z = \frac{x - min(x)}{max(x) - min(x)}$. After normalisation, the values now belong to the interval [0,1].

[1] https://jvion.com
[2] https://www.reddit.com/r/datascience/comments/8c2vnd/what_is_everyones_opinion_on_jvion_and_their/
[3] https://aibrisbane.com.au
[4] https://www.safecareai.com

Data balance: Data imbalance is characterised by a discrepancy in the number of examples per class of a dataset. This phenomenon is known to deteriorate the performance of classifiers, since they are less able to learn the characteristics of the less represented classes [3, 6]. In this way, before using the classifier, data was balanced using SMOTE (Synthetic Minority Over-sampling TEchnique) [1], applied only to training data.

Data imputation: Missing data has been found to have a considerable impact on the learning process of classifiers [7]. Since some fields were missing for some instances in our database, data imputation was performed. Some of the variables were filled in with the value zero, since this value is not in the list for the designation of any attribute. We thus assumed that null meant zero. For the variable "Hour", which only exists in the phase after the operation, the average calculated using only the training data (μ), was used to fill both the training and test data.

Classifiers: The current work deals with binary classification, being the two classes the readmission and no readmission of the patient. Four different classifiers were tested: LR, SVM, DT and kNN.

4 Evaluation

In order to test the performance of the classifiers, the dataset was divided into two sets, a train and a test set. A percentage of 30% was used for the test set, and the remaining was left for the train test. The evaluation methodologies used were the confusion matrix, accuracy, precision, f1-score and recall.

Precision, Recall and F1-Score before surgery are given in Table 1, while the same values for after surgery are summarised in Table 2. Accuracy results are shown in Table 3.

Table 1: Precision, Recall and F1-Score before surgery

Precision	LR	SVM	kNN	DT
Not readmitted	0.64	0.63	0.68	**1.00**
Readmitted	0.85	0.68	0.87	**0.96**
Recall	LR	SVM	kNN	DT
Not readmitted	0.92	0.76	0.92	**0.96**
Readmitted	0.46	0.54	0.54	**1.00**
F1-Score	LR	SVM	kNN	DT
Not readmitted	0.75	0.69	0.78	**0.98**
Readmitted	0.59	0.60	0.67	**0.98**

Table 2: Precision, Recall and F1-Score after surgery

Precision	LR	SVM	kNN	DT
Not readmitted	0.82	0.84	0.80	**0.87**
Readmitted	0.55	0.55	0.65	**0.75**
Recall	LR	SVM	kNN	DT
Not readmitted	0.68	0.76	0.76	**0.86**
Readmitted	0.72	0.73	0.65	**0.75**
F1-Score	LR	SVM	kNN	DT
Not readmitted	0.74	0.80	0.78	**0.86**
Readmitted	0.62	0.67	0.62	**0.75**

Table 3: Accuracy results

Accuracy	LR	SVM	kNN	DT
Before surgery	0.69	0.65	0.73	**0.98**
After surgery	0.69	0.75	0.72	**0.82**

The classifier with the best performance was Decision Trees, achieving an accuracy of 98% for the before surgery experiments and of 82% for after surgery. The worst classifier was SVM with an accuracy of only 65% for before surgery and of 75% after surgery.

Looking at the state of the art, Accuracy is between 69% and 72% in [4], while the Accuracy reported in [5] reaches values between 64% – 70%. It can be thus be concluded that our results exceeded the ones previously published.

For this scenario, it is more important to predict that a patient will be readmitted, even when he ends up not being readmitted. This prediction will allow for preventive measures to be undertaken. When analysing the confusion matrices (not shown due to space constrictions), it could be seen that, before surgery, only one case were miss-classified as "Not readmitted" when in fact they were readmitted within 30 days. For after surgery, 428 cases were miss-classified as "Not readmitted" when in fact they were readmitted.

Having selected the best model, Decision Trees, it is important to access its stability. In this way, a set of 30 runs was performed, each time with different randomly selected train and set sets (always in the proportion of 70%/30%). Average and standard deviation of the accuracy in each run is summarised in Table 4.

Table 4: Accuracy stability assessment

	Average	Standard deviation
Before surgery	0.91	0.03
After surgery	0.82	0.01

It can be seen that accuracy average values are very close to the ones previously stated, and standard deviation are low, assuring the models' stability.

5 Conclusions

The objective of this project was to develop a forecast model for the readmission of a patient before and after undergoing a surgical intervention. The existence of these models will have a high impact in the clinical practice. A patient predicted to be readmitted can be more carefully analysed by the healthcare staff and more tests and procedures can be performed before his release from the hospital in order to reduce the number of readmissions. Even after the release of the patient from the hospital, a closer monitoring of the recovery by phone calls or schedule appointments can be done to identify early possible problems.

Although the developed model is functional, there are improvements that could be made. The application of deep learning techniques [2] is a possibility. We note, however, that these type of models need to be carefully evaluated, being that simpler models are to be favoured in this context.

6 Acknowledgements

This work is partially financed by National Funds through the Portuguese funding agency, FCT - Fundação para a Ciência Tecnologia within project UIDP/00776/2020.

References

[1] NV Chawla, KW Bowyer, LO Hall, and WP Kegelmeyer. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 16:321–357, 2002.

[2] I Domingues and J S Cardoso. Mass detection on mammogram images: a first assessment of deep learning techniques. In *19th Portuguese Conference on Pattern Recognition*, 2013.

[3] I Domingues, JP Amorim, PH Abreu, H Duarte, and J Santos. Evaluation of Oversampling Data Balancing Techniques in the Context of Ordinal Classification. In *IEEE International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, 2018.

[4] HJ Lai, PC Chan, HH Lin, YF Chen, CS Lin, and JC Hsu. A Web-Based Decision Support System for Predicting Readmission of Pneumonia Patients after Discharge. In *IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 2305–2310, 2018.

[5] HJ Lai, TH Tan, CS Lin, YF Chen, and HH Lin. Designing a clinical decision support system to predict readmissions for patients admitted with all-cause conditions. *Journal of Ambient Intelligence and Humanized Computing*, 2020.

[6] F Marques, H Duarte, J Santos, I Domingues, JP Amorim, and PH Abreu. An iterative oversampling approach for ordinal classification. In *34th ACM/SIGAPP Symposium on Applied Computing*, pages 771–774, 2019.

[7] RC Pereira, MS Santos, PP Rodrigues, and PH Abreu. MNAR Imputation with Distributed Healthcare Data. In *EPIA Conference on Artificial Intelligence*, pages 184–195, 2019.

[8] B Sousa-Pinto, AR Gomes, A Oliveira, C Ivo, G Costa, J Ramos, J Silva, MC Carneiro, MJ Domingues, MJ Cunha, A Costa-Pereira, and A Freitas. Reinternamentos hospitalares em Portugal na última década. *Acta Medica Portuguesa*, 2013.

Evaluating a lightweight neural reranking model for biomedical question answering

Tiago Almeida
tiagomeloalmeida@ua.pt
IEETA
Universidade de Aveiro
Aveiro, Portugal

Sérgio Matos
aleixomatos@ua.pt
DETI/IEETA
Universidade de Aveiro
Aveiro, Portugal

Abstract

Automatic searching mechanisms are essential to human progress by simplifying access to relevant information in increasingly large libraries.

In this paper, we present a lightweight searching system, that combines traditional techniques with neural networks yielding a model with only 620 trainable parameters that can be applied to any searching problem for which there is training data available.

We evaluated our system in two challenges, both on the biomedical domain, namely BioASQ 8b and TREC-Covid. In the first one, we achieved top and close to top scores in all the batches, while on TREC-Covid our best result was a second place in the third round.

1 Introduction

In today's science, we witness an unprecedented amount of new information being generated every year. So, the ability to automatically search this unstructured information, like documents, articles, or web pages, becomes a cornerstone of scientific development and progress.

As an example, in the biomedical area, scientists need to routinely search a constantly increasing amount of information, usually in the form of scientific articles, to conduct their day-to-day tasks, which becomes an extremely time-consuming effort. To give a better context, during the current pandemic situation more than 200 thousand articles exclusively related to the study of the coronavirus were published, at a rhythm of approximately one thousand new articles per day[1]. In a more global view, the most used database PubMed/MEDLINE has 30 million indexed articles and is growing at a rate of one and a half million new articles per year.

This searching challenge is addressed by the Information Retrieval (IR) field that studies and creates automatic systems capable of retrieving the most relevant piece of information (usually documents) from a set of unstructured information (set of documents also designated corpus) given a query that encodes the information need. Nowadays, the IR field is considered to be divided into traditional IR and neural IR. The former uses handcrafted rules and equations to directly compute the query-document importance, the BM25 [8] ranking equation being the most popular example. On the other hand, neural IR explores the increasing success of neural networks to approximate a (sub)optimal ranking function by exploring labeled examples. In the literature, the most successful neural architectures for this type of search are Interaction Based, which create a joint representation of the question and the documents by considering multiple matching signals. DRMM [5] and DeepRank [7] are examples of such neural architecture.

This paper presents our lightweight neural interaction-based system, with only 620 trainable parameters, to tackle the previously enunciated searching problem. This system follows a two-step approach where we combined the BM25, a traditional approach, with our lightweight neural model.

We evaluate our system on the 8th BioASQ challenge and on the TREC-Covid challenge, where for the BioASQ we achieved top and close to the top scores for all the batches, while in the TREC-Covid we achieved a second place on the third round as the best result. We are also participated in the TREC-Deep Learning and TREC-Precision Medicine challenges using this same system. However, at the time of writing the results are not available.

[1]These values are inferred from the metadata from the CORD-19 Corpus `https://www.semanticscholar.org/cord19`

2 System Description

As previously mentioned, our system follows a two-step retrieval strategy, more precisely, we adopt the BM25, as the first step, to act as a filter in order to reduce the enormous search space and select only the *top-N* most relevant documents that are further ranked by the neural model, as described in Figure 1.

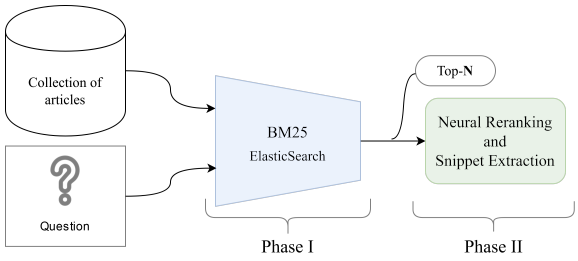


Figure 1: Overview of the system flow.

Our lightweight neural ranking model is a direct evolution of the following previous work [1, 4]. This new version was designed to weight the importance of the document sentences concerning the query by taking into consideration the context where the exact match occurs, *e.g.*, this model produces a more refined judgment of the previously exact match signal considered in the BM25.

Additionally, supported by the inner structure of the designed neural model we devised a zero-shot learning algorithm for sentence extraction, which gave us the ability to extract the sentences that are supposedly more relevant to a given question. This was engineered by looking at the activation value that the model gave to each sentence and retrieving those with higher values. So we assumed that the sentences that most contribute to the final document score must be also the most relevant sentences present in that document for that given question. For a more complete view of the neural model and the zero-shot snippet extraction, we redirect the reader to the following paper [2].

3 Evaluation

In this section, we will individually describe each competition and task following by the respective results.

3.1 BioASQ

The BioASQ challenge [9] is an annual competition on document classification, retrieval, and question-answering, currently in the eighth edition. We submitted our system to be evaluated on the document and snippet retrieval task, part of BioASQ task 8b phase A. For the document retrieval task the objective was to retrieve the most relevant articles from last year's PubMed/MEDLINE annual database. The snippet task is similar but the unit of information becomes the sentences from the PubMed/MEDLINE articles.

The organizers published, in intervals of two weeks, a total of 500 biomedical questions split into five batches. For each batch we submit our system results that were evaluated in terms of Recall, F1, MAP, and GMAP.

With respect to the system, the BM25 filter was fine-tuned with the 2700 biomedical questions provided by the organizers as the training data. The neural model was trained on the same data using a pairwise cross-entropy loss with cyclic learning rates. We also used the GenSim imple-

mentation of the word2vec [6] algorithm to train the word embeddings directly on the PubMed/MEDLINE articles.

Regarding the competition, this edition received on average a total of 25 submissions from 9 teams for the five batches.

Table 1: Summary of the results obtained in the BioASQ.

System	Document Retrieval			Snippet Retrieval		
	Rank	MAP@10	GMAP@10	Rank	MAP@10	F1@10
Batch 1						
Ours	1	33.98	**1.20**	5	29.53	15.00
Top Competitor	3	33.59	0.88	1	85.75	17.52
Batch 2						
Ours	3	31.68	**2.23**	4	27.67	14.13
Top Competitor	1	33.04	1.85	1	68.21	17.73
Batch 3						
Ours	4	43.69	**2.04**	5	41.37	16.61
Top Competitor	1	45.10	1.87	1	100.39	21.40
Batch 4						
Ours	4	40.24	1.31	7	36.59	17.23
Top Competitor	1	41.63	2.04	1	102.44	21.51
Batch 5						
Ours	1	48.42	**3.49**	5	43.79	19.60
Top Competitor	2	48.25	2.54	1	112.67	24.91

In terms of results, we achieved highly competitive scores on the document retrieval task, as shown in Table 1, being first on the first and fifth batches. On the other hand, although our results in the snippet retrieval task were comparatively lower, we consider the results to be encouraging, especially in terms of F1 score, given that these were obtained without supervision.

3.2 TREC-Covid

TREC-Covid was an initiative to rapidly promote the development of an automatic system capable of searching the fast growing literature about the novel coronavirus to aid scientists in their researches. This challenge was organized, in a matter of weeks, by the Allen Institute for Artificial Intelligence (AI2), the National Institute of Standards and Technology (NIST), Oregon Health Science University (OHSU), and others.

The challenge follows a TREC style format and relies on the CORD-19 dataset[2] as the collection of scientific articles about the novel coronavirus. The objective was to retrieve the most relevant articles from this collection for each topic given by the organizers. In TREC challenges, the topic represents the information need that in this case can be used as the query to search the collection.

The competition had a total of five rounds, each with an increasing number of topics: the first round had a total of 30 topics, with increments of 5 topics for the following rounds. The system results were evaluated in a residual manner, except for the first round, since the remaining rounds share topics that had already been evaluated. The metrics adopted were P@5, NDCG@10, Brepf, and MAP. The organizers also allowed the use of the evaluation feedback of previous rounds as training data to tune/train the submitted systems. An important note is that for the first round no training data was available since no previous evaluation had been performed.

Concerning the system, we utilized the BioASQ system on the first round, which means this was a transfer learning approach, exploiting the proximity of the domains. The only change was the embeddings that were trained on the PubMed/MEDLINE articles and the CORD-19 dataset. For the remaining rounds, we kept a similar approach but we also fine-tuned (trained) the model with the feedback data from previous rounds. For a more complete description of the strategy followed, we redirect the reader to the following work [3].

Regarding the competition, this challenge received a lot of interest from the community, resulting in one of the highest TREC participation rates ever. For example, in the first round, a total of 56 teams submitted results for a total of 143 runs.

Table 2: Summary of the two best results achieved on TREC-Covid.

System	Round 1			Round 3		
	Rank	P@5	NCDG@10	Rank	P@5	NCDG@10
Ours	9	63.33	52.98	2	**86.50**	77.15
Top Competitor	1	78.00	60.80	1	86.00	77.40

[2] https://www.semanticscholar.org/cord19

In terms of results, we achieved positive and encouraging results, being the best one on the third round as shown in Table 2. Furthermore, we show that our assumption to perform transfer learning with the BioASQ data empirically works, by beating traditional IR techniques and more recent transform-based techniques like BERT and T5. Regarding the remaining rounds, we scored approximately in the middle of the table. This lower results could be partially explained by mistakes later identified in these submissions.

4 Conclusion

In this paper, we show a two-stage retrieval system that was evaluated on two biomedical challenges, namely BioASQ 8b and TREC-Covid. Regarding the BioASQ 8b, we demonstrate the effectiveness of our system on the document task, by achieving top scores, and show a promising zero-shot learning setup for the snippet retrieval task. With respect to the TREC-Covid challenge, we demonstrate a successful transfer learning technique of our BioASQ system to this new task by leveraging the proximity of domains between the tasks.

Acknowledgements

This work was supported by National Funds through the FCT - Foundation for Science and Technology, in the context of the project UIDB/00127/2020, and by the EU/EFPIA Innovative Medicines Initiative 2 Joint Undertaking undergrant agreement No 806968.

References

[1] Tiago Almeida and Sérgio Matos. Neural-based snippet extraction for biomedical question answering. *Proceedings of the 25th Portuguese Conference on Pattern Recognition*, 2019. URL http://recpad2019.dcc.fc.up.pt/wp-content/uploads/2019/05/ProceedingsRECPAD.pdf#page=79.

[2] Tiago Almeida and Sérgio Matos. BIT.UA at BioASQ 8: Lightweight neural document ranking with zero-shot snippet retrieval. *BioASQ 8 workshop, CLEF 2020*, 2020.

[3] Tiago Almeida and Sérgio Matos. Frugal neural reranking: evaluation on the covid-19 literature. 2020. URL https://openreview.net/pdf?id=TtcUlbEHkum.

[4] Tiago Almeida and Sérgio Matos. Calling attention to passages for biomedical question answering. In *Advances in Information Retrieval*, pages 69–77, Cham, 2020. Springer International Publishing. ISBN 978-3-030-45442-5. doi: 10.1007/978-3-030-45442-5_9.

[5] Ryan McDonald, Georgios-Ioannis Brokos, and Ion Androutsopoulos. Deep Relevance Ranking Using Enhanced Document-Query Interactions. sep 2018. URL http://arxiv.org/abs/1809.01682.

[6] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, page 3111–3119, Red Hook, NY, USA, 2013. Curran Associates Inc.

[7] Liang Pang, Yanyan Lan, Jiafeng Guo, Jun Xu, Jingfang Xu, and Xueqi Cheng. Deeprank. *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, Nov 2017. doi: 10.1145/3132847.3132914.

[8] Stephen Robertson and Hugo Zaragoza. The probabilistic relevance framework: Bm25 and beyond. *Found. Trends Inf. Retr.*, 3(4):333–389, April 2009. ISSN 1554-0669. doi: 10.1561/1500000019.

[9] George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael Alvers, Dirk Weißenborn, Anastasia Krithara, Sergios Petridis, Dimitris Polychronopoulos, Yannis Almirantis, John Pavlopoulos, Nicolas Baskiotis, Patrick Gallinari, Thierry Artieres, Axel-Cyrille Ngonga Ngomo, Norman Heino, Eric Gaussier, Liliana Barrio-Alvers, and Georgios Paliouras. An overview of the bioasq large-scale biomedical semantic indexing and question answering competition. *BMC Bioinformatics*, 16:138, 04 2015. doi: 10.1186/s12859-015-0564-6.

Brain Extraction for Analysis of Magnetic Resonance Imaging in Patients with Multiple Sclerosis

Marcela de Oliveira
marcela.oliveira@unesp.br
School of Sciences, São Paulo State University, UNESP - Brazil

Marina Piacenti-Silva
School of Sciences, São Paulo State University, UNESP - Brazil

Paulo Noronha Lisboa-Filho
School of Sciences, São Paulo State University, UNESP - Brazil

Fernando Coronetti Gomes Rocha
Medical School, São Paulo State University, UNESP - Brazil

Jorge Manuel Santos
ISEP, School of Engineering, Polytechnic of Porto - Portugal

Jaime dos Santos Cardoso
INESC TEC and Faculty of Engineering, University of Porto – Portugal

Abstract

Multiple sclerosis (MS) is a neurodegenerative disease that is increasing worldwide. MS diagnosis and monitoring treatment is vitally important. Due to neuronal damage, which occurs in neurons, this disease affects the ability of nerve cells in the brain and spinal cord to communicate. Magnetic resonance imaging is the gold standard exam for diagnosis and monitoring of multiple sclerosis. MS is characterized by brain lesions where the neurodegeneration process occurs, making it possible to visualize these affected areas on magnetic resonance images. The advancement of technology has allowed an improvement in the sequences for the visual detection of lesions caused by multiple sclerosis, aiding medical diagnosis. Imaging pre-processing is an important step in analysing specific structures. Thus, the purpose of this work is to perform image processing in MRI with skull stripping from MS patients for future detection and quantification of brain lesions. We concluded that the automatic pre-processing method applied in this work for skull stripping can be used for the brain extraction process and for future sclerotic lesions identification.

1 Introduction

Magnetic resonance imaging (MRI), due to the richness in the information details provided, is the gold standard exam for diagnosis and follow-up of neurodegenerative diseases, such as multiple sclerosis (MS) [1]. There is increasing prevalence and incidence of MS in both developing and developed countries [2, 3]. MS is a chronic neurological disease characterized by demyelination of axons [4]. This demyelination process (neurodegeneration) causes lesions in white matter that can be observed in vivo by MRI. In individuals with MS, radiological abnormalities can be identified even in the absence of clinical symptoms of the disease, and the areas where demyelination occurs can be seen in this type of image [5]. MRI allows the evaluation and follow-up of sclerotic lesions in different sequences such as T1, T2 and FLAIR (Fluid Attenuated Inversion Recovery) [6].

The sclerotic lesions observed in the T1-weighted MRI sequence are areas with less signal intensity when compared to normal areas [7]. In this type of sequence, the injured area becomes isointense within a few months after the cessation of inflammatory activity and with the process of repairing mechanisms, such as remyelination. The highlight of the lesions can also be seen in the T2 and FLAIR sequences, the affected area is characterized by hypersignal. Such lesions may provide quantitative assessments of the inflammatory activity of the disease, and possibly heralding future brain atrophy and clinical disability. Quantitative measures based on various features of lesions have been shown to be useful in clinical trials for evaluating therapies. In order to perform the identification and quantification of sclerotic lesions, it is necessary to perform a pre-processing of the images to extract the brain [8]. Thus, the purpose of this work is to perform image processing in MRI with skull stripping from MS patients for future detection and quantification of brain lesions.

2 Materials and Methods

2.1 Patient Sample

Patients in this test group were diagnosed with MS according to the McDonald criteria [9, 10], and recruited from the Hospital of Clinics Botucatu-Brazil (HCB). The dataset is not public and includes a test group which have 5 subjects with 10 scans. Each subject includes two types of sequences: T1 weighted (-w) and FLAIR. All datasets were fully anonymized for dissemination purposes. All the patients' imaging examinations and diagnostic evaluations of the test group were retrospectively obtained between 2014 and 2019. Patient information was acquired and analyzed in accordance with ethical committees of the author´s institutions, and all the patients gave their written consent to participate in the study.

2.2 Imaging Processing

For all MRI scans, the dataset contained the same number of images in T1-w and FLAIR sequences. MRIs were preprocessed in three steps: 1. rigidly registered; 2. skullstripped; and 3. corrected for intensity inhomogeneity [8]. In the first step the T1-w MRI of subjects in the test group were rigidly registered to the axial 1 mm^3 through general registration (BRAINS) [11]. The FLAIR images (other sequence) were registered to the T1-w image space, by applying the registration transform to the initial volume FLAIR, we generate a new volume spatially aligned with the volume T1. A first step example is shown in Figure 1, where an original T1-w image was rigidly registered for 1mm^3, and FLAIR to T1 space registration.

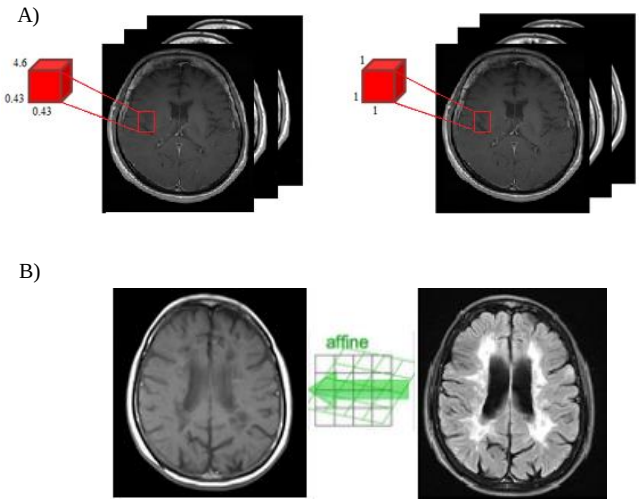


Figure 1: First step example. A) Original image voxel size 0.43x0.43x4.6 mm^3 and image rigidly registered to 1 mm^3. B) FLAIR to T1 space registration.

In the second step, both sequences were stripped by swiss skull stripper [12]. At this point, the algorithm registered a grayscale atlas image to the grayscale patient data. Through registration transform, an atlas mask was propagated with patient data. This brain mask was eroded and served as initialization for a refined brain extraction. Finally, in order to correct the nonuniform intensity in magnetic resonance images caused by field inhomogeneities, the third step performed image bias correction by N4ITK after brain stripping [13].

3 Results

The imaging preprocessing was performed using MRI with T1-w and FLAIR sequences. The results for the first step of the processing was the rigidly register, where the size images 0.43x0.43x4.6mm^3 were transformed to 1mm^3 (see Figure 1.A). The results of special registration from FLAIR to T1 is represented in the Figure 1.B. Figure 2 shows the skullstripping (second step) and bias correction (third step) processes for brain extraction. Brain extraction process was applied to all slices of the exam, and we obtained the brain volume (see Figure 3).

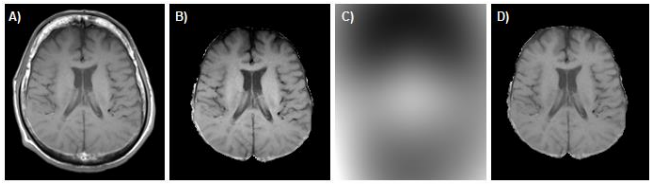


Figure 2: A) Rigidly registered image. B) Skull stripped image. C) Bias correction. D) Final image.

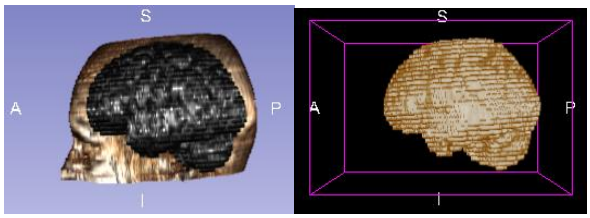


Figure 3: Brain Volume after brain extraction process applied to all slices.

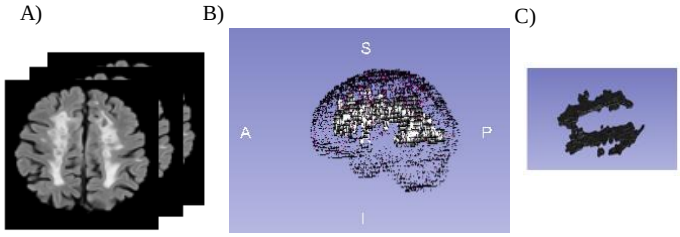


Figure 4: A) Images for identification and segmentation of sclerotic lesions. B) Brain with lesions volumetric representation. C) Segmented volumetric lesion.

4 Conclusions and Future Work

Radiologically, areas where demyelination may occur can be observed by magnetic resonance imaging. We concluded that the automatic preprocessing method applied in this work for skull stripping can be used for the brain extraction process. This is an important and necessary pre-process for future analysis of brain lesions. Thus, the development and application of computer programs can contribute to assist health professionals in the diagnosis and monitoring of patients with neurodegenerative diseases. Manual segmentations brain lesions in MRIs is considered as the gold standard, however, this process is time consuming (the lesion has to be manually segmented in each slice) and suffers intra- and inter-observer variability [14]. Figure 4 shows an example where the segmentation was performed on all slices and represented volumetrically. In future works, we expect to implement more automated methods to lesions identification and segmentation process, including machine-learning approaches, to provide accurate analysis. In addition, we will perform the automatic lesion quantification to assist physicians to decide whether they should follow a treatment with a disease modifying therapy modality, as well as identifying the disease progression.

Acknowledgment

Authors thank the support of Hospital of Clinics- Botucatu Medical School (HC-FMB) of São Paulo State University (UNESP), Botucatu Campus, School of Sciences of São Paulo State University, Bauru Campus, both from Brazil; and the support of INESC TEC - Institute for Systems and Computer Engineering, Technology and Science and FEUP - Faculty of Engineering, University of Porto, Portugal. This work was supported by a grant from Brazilian agency Fundação de Amparo à Pesquisa do Estado de São Paulo (number 2019/16362-5 and 2017/20032-5).

References

[1] X. Lladó, O. Ganiler, A. Oliver, R. Martí, J. Freixenet, L. Valls, J. C. Vilanova, L. Ramió-Torrentà, A. Rovir. Automated detection of multiple sclerosis lesions in serial brain MRI. *Neuroradiology*, 54(8):787-807, 2012.

[2] P. Browne, D. Chandraratna, C. Angood, H. Tremlett, C. Baker, B. V. Taylor, A. J. Thompson. Atlas of Multiple Sclerosis 2013: A growing global problem with widespread inequity. *Neurology*, 83(11):1022-1024, 2014.

[3] R. Dobson, G. Giovannoni. Multiple sclerosis – a review. *European Journal of Neurology*, 26(1):27-40, 2019.

[4] B. Derkus, E. Emregul, C. Yucesan, K.C. Emregul. Myelin basic protein immunosensor for multiple sclerosis detection based upon label-free electrochemical impedance spectroscopy. *Biosensors and Bioelectronics*, 46:53-60, 2013.

[5] S.V. Ramagopalan, R. Dobson, U. C. Meier, G. Giovannoni. Multiple sclerosis: risk factors, prodromes, and potential causal pathways. *The Lancet Neurology*, 9(7):727-739, 2010.

[6] R. Bakshi, A. J. Thompson, M. A. Rocca, D. Pelletier, V. Dousset, F. Barkhof, M. Inglese, C. R. Guttmann, M. A. Horsfield, M. Filippi. MRI in multiple sclerosis: current status and future prospects. *The Lancet Neurology*, 7(7):615-625, 2008.

[7 J. F. Kurtzke. Rating neurologic impairment in multiple sclerosis an expanded disability status scale (EDSS). *Neurology*, 33(11):1444-1444, 1983.

[8] M. de Oliveira, M. Piacenti-Silva, F. C. G. Rocha, J. M. Santos, J. S. Cardoso, P. N. Lisboa-Filho. Skull Extraction for Quantification of Brain Volume in Magnetic Resonance Imaging of Multiple Sclerosis Patients. In: *ICMPBE 2020 : International Conference on Medical Physics and Biomedical Engineering: July 27-28, 2020; Zurich, Switzerland*: World Academy of Science, Engineering and Technology - Biomedical and Biological Engineering, 14(7), 2020.

[9] W. I. McDonald, A. Compston, G. Edan, D. Goodkin, H. P. Hartung, F. D. Lublin, H. F. McFarland, D. W. Paty, C. H. Polman, S. C. Reingold. Recommended diagnostic criteria for multiple sclerosis: guidelines from the International Panel on the Diagnosis of Multiple Sclerosis. *Annals of neurology*, 50, 2001.

[10] A. J. Thompson, B. L. Banwell, F. Barkhof, W. M. Carroll, T. Coetzee, G. Comi, J. Correale, F. Fazekas, M. Filippi, M. S. Freedman *et al*. Diagnosis of multiple sclerosis: 2017 revisions of the McDonald criteria. *The Lancet Neurology*, 17(2):162-173, 2018.

[11] H. J. Johnson, G. Harris, K. Williams. BRAINSFit: Mutual Information Registrations of Whole-Brain 3D Images, Using the Insight Toolkit. *The Insight Journal*, 2007.

[12] S. Bauer, T. Fejes, M. Reyes. A Skull-Stripping Filter for ITK. *The Insight Journal*, 2013.

[13] N. J. Tustison, B. B. Avants, P. A. Cook, Y. Zheng, A. Egan, P. A. Yushkevich, J. C. Gee. N4ITK: improved N3 bias correction. *IEEE Trans Med Imaging*, 29(6):1310-1320, 2010.

[14] S. Jain, D. M. Sima, A. Ribbens, M. Cambron, A. Maertens, W. Van Hecke, J. De Mey, F. Barkhof, M. D. Steenwijk, M. Daams *et al*. Automatic segmentation and volumetry of multiple sclerosis brain lesions from MR images. *NeuroImage Clinical*, 8:367-375, 2015.

Adversarial learning for a robust fingerprint presentation attack detection method against unseen attacks

João Afonso Pereira[1]
joao.p.pereira@inesctec.pt
Diogo Pernes[1,3]
dpc@inesctec.pt
Ana F. Sequeira[1]
ana.f.sequeira@inesctec.pt
Jaime S. Cardoso[1,2]
jaime.cardoso@inesctec.pt

[1] INESC TEC
Porto, Portugal
[2] Faculdade de Engenharia da Universidade do Porto
Porto, Portugal
[3] Faculdade de Ciências da Universidade do Porto
Porto, Portugal

Abstract

Fingerprint presentation attack detection (PAD) methods present a stunning performance in current literature. However, the *fingerprint PAD generalisation problem* is still an open challenge requiring the development of methods able to cope with sophisticated and unseen attacks as our eventual intruders become more capable. This work addresses this problem by applying a regularisation technique based on an adversarial training and representation learning specifically designed to improve the PAD generalisation capacity of the model to an unseen attack. The application of the adversarial training methodology is evaluated in two different scenarios: i) a handcrafted feature extraction method combined with a Multilayer Perceptron (MLP); and ii) an end-to-end solution using a Convolutional Neural Network (CNN). The experimental results demonstrated that the adopted regularisation strategies equipped the neural networks with increased PAD robustness. The CNN models' capacity for attacks detection in the unseen-attack scenario was particularly improved, showing remarkable improved APCER error rates when compared to state-of-the-art methods in similar conditions.

1 Introduction

Fingerprint presentation attack detection (FPAD) methods have been developed to overcome the vulnerability of fingerprint recognition systems (FRS) to spoofing. However, most of the traditional approaches have been quite optimistic about the behavior of the intruder, assuming the use of a previously known type of attack sample. This assumption has led to the overestimation of the performance of the methods, using both live and spoof samples to train the predictive models and evaluate each type of fake samples individually [10].

In this work, the *FPAD generalisation problem* is addressed by means of a regularisation technique designed to improve the generalisation capacity to unseen attacks in which the proposed model jointly learns the representation and the classifier from the data, while explicitly imposing invariance to the presentation attack instrument (PAI) types aka, 'PAI-species', in the high-level representations for a robust PAD method. The contributions of this work are then two-fold: 1) application of the adversarial training concept to the generalisation to unseen attacks problem in FPAD; and 2) evaluation of the adversarial training methodology in: i) combination of handcrafted features with a Multilayer Perceptron (MLP); ii) a Convolutional Neural Network (CNN) end-to-end solution. In this paper, this section summarises the proposed work and how it addresses the research question posed, section 2 presents the methodology, section 3 describes the experimental setup, section 4 presents the results and discussion and finally section 5 concludes the work.

2 Proposed Methodology

This work applies the methodology from Ferreira *et al*. [1] which was adopted in Pereira *et al*. [9] with the appropriate adjustments. The original method was presented by Ferreira *et al*. [2] in the context of sign language recognition, in an approach that builds on those initially introduced by Ganin *et al*. [4], for domain adaptation, and Feutry *et al*. [3], to learn anonymized representations. The underlying idea behind this approach is that, in order to generalise well to unseen attacks, the model should not specialize in discriminating any of the PAI species (PAISp) presented at training time and, therefore, the learned internal representations should be invariant to the PAISp. For this purpose, the model combines an adversarial approach with a species-transfer training objective The high-level architecture of the model is summarized in Fig. 1. It should be assumed that one has access to a labeled dataset $\mathbb{X} = \{\boldsymbol{X}_i, y_i, s_i\}_{i=1}^{N}$ of N samples, where $\boldsymbol{X}_i$ represents the i-th input sample, and y_i and s_i denote the corresponding class label (*bona fide* or *attack*) and the PAI species (only defined for attack samples), respectively. Let $\mathbb{X}^{bf}$ and $\mathbb{X}^{a}$ be these partitions of $\mathbb{X}$ for bona-fide and attack samples, respectively, and N^{bf} and N^{a} their respective cardinality.

The model comprises three main sub-networks: (i) an encoder network $h(\cdot;\theta_h)$ that receives input samples and maps them to a latent space; (ii) a *task-classifier* network $f(\cdot;\theta_f)$ which aims to distinguish attack and bona fide samples, mapping latent representations to the corresponding class probabilities; and (iii) a *species-classifier* network $g(\cdot;\theta_g)$ that receives latent representations from attack samples and aims to predict the corresponding PAI species. The species-classifier is trained to minimize the classification loss of the PAI-species. Simultaneously, the task-classifier and the encoder are jointly trained to minimize the classification loss between attacks and bona fide samples, while trying to keep the PAI-species classification close to random guessing. In addition to the adversarial training, a species-transfer objective is employed to further encourage the latent representations to be species-invariant. The overall objective function of the encoder and task classifier is then the combination of the previous objectives and can be formulated as:

$$\min_{\theta_h,\theta_f} \mathcal{L}(\theta_h,\theta_f,\theta_g) = \min_{\theta_h,\theta_f} \left\{ \mathcal{L}_{\text{task}}(\theta_h,\theta_f) + \lambda \mathcal{L}_{\text{adv}}(\theta_h,\theta_g) + \gamma \mathcal{L}_{\text{transfer}}(\theta_h) \right\}, \tag{1}$$

where $\gamma \geq 0$ is the weight that controls the relative importance of the species-transfer term and the objective for the species-classifier remains unchanged.

3 Experimental Setup

For more details, the reader is referred to Pereira *et al*. [9].

PAD Performance Evaluation Metrics: *Equal Error Rate (EER)*, *Attack Presentation Classification Error Rate* (APCER) and*Bona-fide Presentation Classification Error Rate* (BPCER) for APCER of 5% as in [6].

Dataset: The Fingerprint LivDet2015 [7] training dataset comprises a set of five subsets, each one corresponding to a specific fingerprint sensor. For each sensor there are bona fide samples and different types of PAI.

Evaluation protocols: The framework is denominated "unseen-attack", as the PAI seen in the testing phase is unknown to the model.

Handcrafted feature extraction method: Histograms of intensity, Local Binary Patterns (LBP and Local Phase Quantization .

Implementation details: The models were implemented in Python with the PyTorch library. For details, see Pereira *et al*. [9].

4 Results and discussion

In Table 2, the results of the baseline methods (*MLP* and *CNN*) and their respective regularised versions (MLP_{reg} and CNN_{reg}) are displayed. Comparing the performance of the baseline and regularised versions, it can be observed that: i) regarding the MLP, except for the Hi Scan sensor, in all the cases there is a significant improvement in at least 2 out of the 3 presented metrics; and ii) regarding the CNN, there is a significant improvement without exception in all error rates, with a particular significant improvement of the APCER value from 4.12% to 0.81% (for the average of the five sensors). From these observations, it can be stated with confidence that, overall, the regularisation technique improves the PAD robustness of both the models.

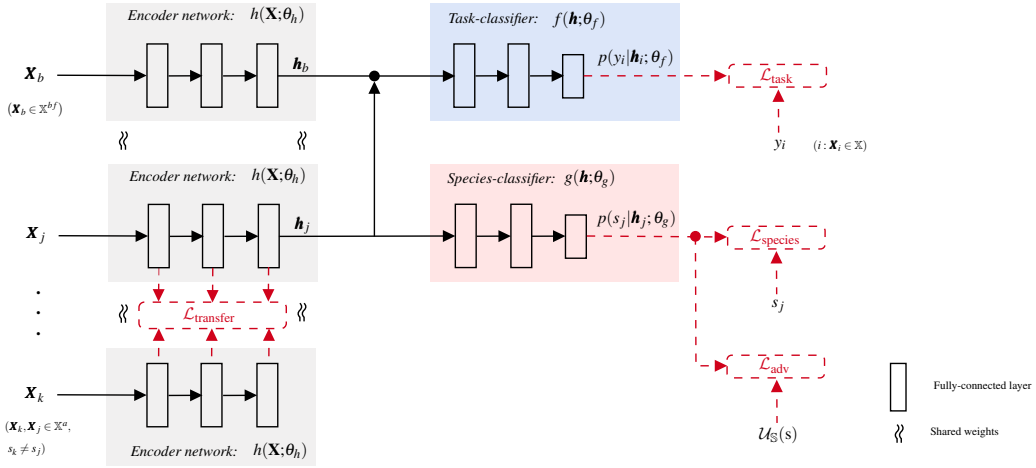


Figure 1: The architecture of the proposed species-invariant neural network (from [1]).

Still, it is arguable that the performance of the MLP, even the baseline version, outperforms the CNN results. Nevertheless, it should be noted that: i) the first scenario is taking advantage of rich handcrafted features; and ii) the data available for training is not enough to take the best out of the CNN learning capabilities. Thus, on the one hand the end-to-end solution provided by the CNN saves a considerable effort in the computation of the feature extraction step and, on the other hand, increasing the amount of training data will certainly increase the performance of these models, as there is a high potential for growth.

Table 1: Baseline and proposed regularised approaches - Cross Match, Digital Persona and Green Bit sensors. (*BPCER@APCER* = 5% noted by *BPCER@5*.)

Method	PAD metrics (%)								
	Cross Match			*Digital Persona*			*GreenBit*		
	APCER	BPCER@5	EER	APCER	BPCER@5	EER	APCER	BPCER@5	EER
MLP	**0.07**	7.57	4.33	0.00	0.53	0.45	**0.70**	**0.20**	1.10
MLPreg	0.13	**4.30**	**3.70**	**0.00**	**0.00**	**0.30**	**0.70**	0.63	**0.93**
CNN	5.00	6.25	8.70	5.60	10.80	7.28	3.03	14.13	7.05
CNNreg	**1.07**	**4.65**	**2.82**	**0.60**	**3.85**	**2.45**	**0.60**	**2.93**	**1.63**

Table 2: Baseline and proposed regularised approaches - Hi Scan and Time Series sensors, as well as the average of the results for the 5 sensors. (*BPCER@APCER* = 5% noted by *BPCER@5*.)

Method	PAD metrics (%)								
	Hi Scan			*Time Series*			Average of the 5 sensors		
	APCER	BPCER@5	EER	APCER	BPCER@5	EER	APCER	BPCER@5	EER
MLP	**0.30**	**2.83**	**3.03**	**0.00**	**0.03**	0.60	**0.21**	2.23	1.90
MLPreg	1.30	3.60	3.38	**0.00**	**0.03**	**0.10**	0.43	**1.71**	**1.68**
CNN	5.60	20.15	11.25	1.37	9.10	4.07	4.12	12.09	7.67
CNNreg	**1.20**	**1.21**	**1.04**	**0.60**	**6.30**	**2.70**	**0.81**	**3.79**	**2.13**

Despite the evidences showed in favour of the effectiveness of the regularisation technique, it is crucial to compare the results obtained with the proposed approach against the current state-of-the-art DL based PAD that tackle the unseen-attack scenario. This is not an easy task as most works still opt for a more traditional approach, based on binary classification limited to one type of attack at a time. From the available literature using similar databases and addressing the generalisation problem, stands out the meritory initiative of Fingerprint LivDet2015 [7] of evaluating the methods with some unseen types of PAISp.

Table 3 presents the results of the proposed regularised CNN version, *CNNreg*, alongside with the comparable literature methods currently available. The comparison shows the best results for common subsets of the used database presented in the LivDet2015 [5, 7] competition, as well as with an additional recent publication [8]. From the observed results, it is remarked the significant improvement of the *CNNreg* in two out of three sensors and undoubtedly when considering the average values. In particular, the *CNNreg* provided an APCER value of 0.76% against 2.09% and 6.33% of the other methods (for the average of the three sensors).

Table 3: Literature and proposed approach.(*BPCER@APCER* = 5% noted by *BPCER@5*.)

Method	PAD metrics (%)							
	Cross Match		*Digital Persona*		*GreenBit*		Average	
	APCER	BPCER@5	APCER	BPCER@5	APCER	BPCER@5	APCER	BPCER@5
Proposed CNNreg	1.07	4.65	**0.60**	**3.85**	**0.60**	**2.93**	**0.76**	**3.81**
LivDet2015 [5, 7]	1.68	≈ **0.80**	**0.60**	≈ 10.00	4.00	≈ 5.00	2.09	≈ 5.27
Park et al. [8]	**0.00**	-	11.00	-	8.00	-	6.33	-

5 Conclusions and future work

Comparing the baseline and regularised versions, it can be stated that, overall, the regularisation technique improves the PAD robustness of both the models. Despite the fact that the *MLPreg* fed with rich handcrafted features proved to be competitive, the fact is that *CNNreg* has more potential for growth and for increasing its performance in the future. The comparison of the proposed approach against the current DL based PAD methods that tackle the unseen-attack scenario, is not an easy task as most works still opt for a more traditional approach based on binary classification limited to one type of attack at a time. Still, from the comparison with the available literature using similar databases and addressing the generalisation problem, it is verified a significant superiority of the *CNNreg* in two out of three sensors and undoubtedly when considering the average values.

Acknowledgements

This work is financed by National Funds through the Portuguese funding agency, FCT - Fundação para a Ciência e a Tecnologia, within project UIDB/50014/2020 and by Ph.D. Grant SFRH/BD/129600/2017.

References

[1] Pedro Ferreira, Ana F. Sequeira, Diogo Pernes, Ana Rebelo, and Jaime S. Cardoso. Adversarial learning for a robust iris presentation attack detection method against unseen attack presentations. In *Proceedings of the 18th BIOSIG*, 2019.

[2] Pedro M. Ferreira, Diogo Pernes, Ana Rebelo, and Jaime S. Cardoso. Learning signer invariant representations with adversarial training. In *12th ICMV*, 2019.

[3] Clément Feutry, Pablo Piantanida, Yoshua Bengio, and Pierre Duhamel. Learning anonymized representations with adversarial neural networks. *arXiv:1802.09386*, 2018.

[4] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *Proc. 32nd Int. Conf. ML*, 2015.

[5] Luca Ghiani, David A. Yambay, Valerio Mura, Gian Luca Marcialis, Fabio Roli, and Stephanie A. Schuckers. Review of the fingerprint livdet competition: 2009 to 2015. *IMAV*, 58:110 – 128, 2017.

[6] ISO/IEC JTC1 SC37. Information Technology - Biometrics - Presentation attack detection Part 3: Testing and Reporting. 2017.

[7] Valerio Mura, Luca Ghiani, Gian Marcialis, Fabio Roli, David Yambay, Schuckers, and Stephanie Schuckers. Fingerprint LivDet2015.

[8] E. Park, X. Cui, T. H. B. Nguyen, and H. Kim. Presentation attack detection using a tiny fully convolutional network. *IEEE TIFS*, 14 (11):3016–3025, 2019.

[9] Joao Pereira, Ana F. Sequeira, Diogo Pernes, and Jaime S. Cardoso. A robust fingerprint presentation attack detection method against unseen attacks through adversarial learning. In *19th BIOSIG*, 2020.

[10] Ana F. Sequeira and Jaime S. Cardoso. Fingerprint liveness det. in the presence of capable intruders. *Sensors*, 15:14615–14638, 2015.

Removal of periodic geometric structure in the fingerprint minutiae detection

Eduardo Castro[1]
eduardo.m.castro@inesctec.pt
Ana Rebelo[1]
arebelo@inesctec.pt
Carlos Gonçalves[2]
Carlos.Goncalves@incm.pt
Jaime S. Cardoso[1]
jaime.cardoso@inesctec.pt

[1] INESC TEC
Rua Dr. Roberto Frias
Porto, Portugal

[2] INCM
Edifício Casa da Moeda
Av. António José de Almeida
Lisboa, Portugal

Abstract

The main feature of the Portuguese Citizen Card is to allow the civil identification, in person or at a distance using electronic devices. The biometric identification is done through fingerprint images using specific points, called minutiae, for the matching. In this paper, a method to remove periodic geometric structure in the detected minutiae is proposed. The aim is to improve the interoperability according to the Minutiae Interoperability Exchange (MINEX) III program.

1 Introduction

An increasing number of biometrics have been deployed in real-world applications and its use is becoming a daily life practice for an ever growing number of people around the world. In consequence, a high level of reliability and robustness is required for sensitive applications such as border control, access control to military or laboratory facilities, as well as access to personal accounts for mobile on-line banking. Biometric traits can provide this automatic recognition measuring unique physical or behavioral characteristics.

Notwithstanding face has been preferred in a number of biometric applications (such as border control e-gates, on-line banking apps, CCTV surveillance identification, selfie-based authentication on smartphones, among others), fingerprint is still one of the most used biometric traits principally because of its social acceptance and stability.

Portugal was the pioneer with the "Match-on-Card" (MoC) fingerprint matching algorithm implemented in the national eID card. This technique brought very significant changes in this state: 1) modernization, 2) simplification and 3) technical evolution. The Portuguese National Printing Office – INCM (Imprensa Nacional Casa da Moeda SA), responsible for the creation of the method, provided to the Portuguese Government an innovative way of fingerprint matching in the card microprocessor without any contact to a central biometric database. The biometric information and the technology inserted in the Citizen Card allows an high security authentication. In this context, a national fingerprint recognition algorithm, hereafter referred to as fingerIDAlg, capable of MoC was developed by a Portuguese R&D institute in partnership with INCM. The main contributions of this work were: 1) an algorithm with higher accuracy than the previous solution; 2) an extremely competitive time processing MoC algorithm; and 3) an independent proprietary sensor solution [6].

Nevertheless, the proposed solution could still be improved in terms of minutiae geometric structure in order to be in compliance with NIST's Minutiae Interoperability Exchange (MINEX) III criteria guidelines[1]. In this work, a simple but effective solution is presented. The obtained results using the minutiae density plots and from the NIST's report proves exactly that. For architecture details of the fingerprint minutiae extraction algorithm the reader should consult the following paper [6].

2 Minutia Density Plots

Minutia density plots show where the template generator tends to find minutia in fingerprint images. They are 2D histograms where the degree of illumination at an (x, y) coordinate indicates how frequently the software located a minutiae point at that location – see Figure 1. The purpose of showing minutia density plots is to determine whether the template generator exhibits regional preference when locating minutia. Periodic structures and other regional preferences affects interoperability [7].

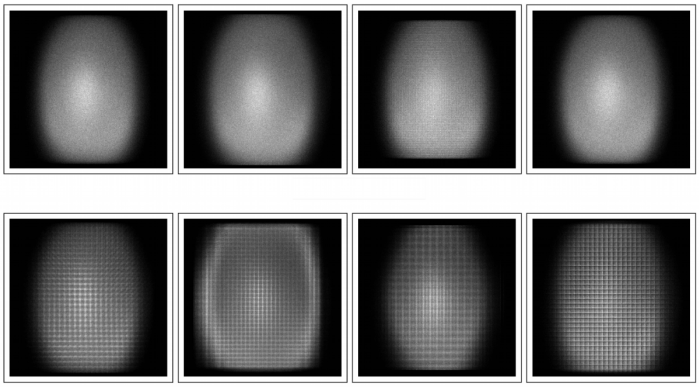


Figure 1: Comparison between natural (top) and unnatural (bottom) structures in the minutiae position.

The criterion regarding the obligation that the algorithm does not exhibit periodic behavior with respect to extracting the minutiae position is evaluated in a quantitative way, through a routine that is published [2], in a set of about 600k fingerprint images. The periodicity is measured as the highest coefficient of the Fourier representation of the minutiae position histogram, after removing some low frequency components. The measured value must be less than to a experimental value. A 0.002 was selected empirically by NIST.

3 Baseline

The positioning of minutiae extracted by fingerIDAlg has a periodic pattern in a fixed size grid – see Figure 2.

The reasoning of this periodic structure is in the way how the fingerIDAlg computes the orientation for each pixel in the image. In this module, an angle is calculated in an window of K pixels and the remaining pixels are obtained by interpolation. In this way the processing time is very significantly reduced and the use of local orientations allows less noise in the angles of the detected minutiae which an important factor in the matching process.

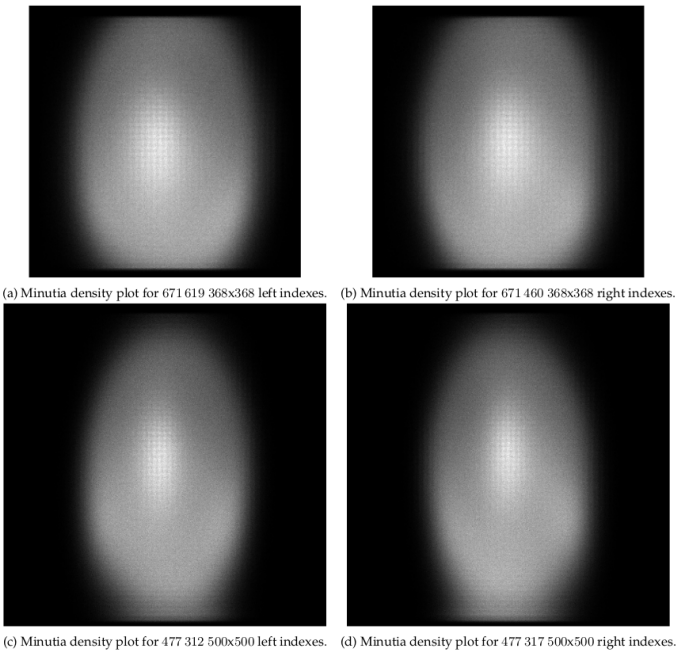
4 The Proposed Solution

The structure presented by fingerIDAlg in the extraction of minutiae corresponds to a periodic pattern in a fixed size grid determined by parameter K. In this manner, initially the image is moved X pixels to the right and Y pixels down. X and Y are determined in a pseudo-random manner in order to guarantee: 1) for the same input the same values are used; 2) the values vary between 0 and K-1 and 3) the possible values have the same probability of being obtained. The image is then processed by fingerIDAlg to extract the minutiae – see Figure 3. At the end, the position of each extracted minutia is corrected, moving again X pixels to the left and Y pixels upwards.

The proposed approach intends to *dilute* the grid previously presented in the histogram. Points with a high probability of generating minutiae

[1] https://www.nist.gov/itl/iad/image-group/minex-iii-compliance-guidelinesvisitadadia1/02/2020.

[2] https://github.com/usnistgov/minex/tree/master/minexiii/grid_detector



(a) Minutia density plot for 671 619 368x368 left indexes. (b) Minutia density plot for 671 460 368x368 right indexes.

(c) Minutia density plot for 477 312 500x500 left indexes. (d) Minutia density plot for 477 317 500x500 right indexes.

Figure 2: The obtained results of fingerIDAlg in the NIST's MINEX III.

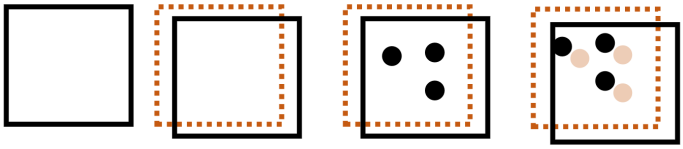


Figure 3: The proposed methodology to remove the periodic structure from the minutiae extraction phase.

will be distributed across the pixels in a $K \times K$ size square. Given its computational simplicity and the fact that it is a pre-and-pro-processing methodology the influence in the final accuracy and processing time is small.

5 Experimental Evaluation

5.1 Datasets

The Synthetic FINgerprint GEnerator (SFinGe) [5] was used to test offline the proposed methodology. In total, 800k images with a dimension of 260×264 and a resolution of 500 dpi were generated in SFinGe. Fingerprint images with background noise, pores, scars and cuts were generated. Different orientations were also included. Experiments were also conducted using the FVC databases: FVC2000 [3], FVC2002 [4], FVC2004 [1] and FVC2006 [2]. Each FVC database is composed of 4 subsets (DB1 A, DB2 A, DB3 A and DB4 A). The first 3 sets have a total of 800 images, acquired from 100 fingers with 8 samples per finger. FVC2006 comprises 1680 fingerprints images acquired from 140 fingers with 12 samples per finger. In total, 12240 fingerprints are available for testing the algorithms. The images have a resolution ranging from 250 to 569 dpi. The dimensions vary from 96 to 640 pixels in width, and 96 to 480 pixels in height.

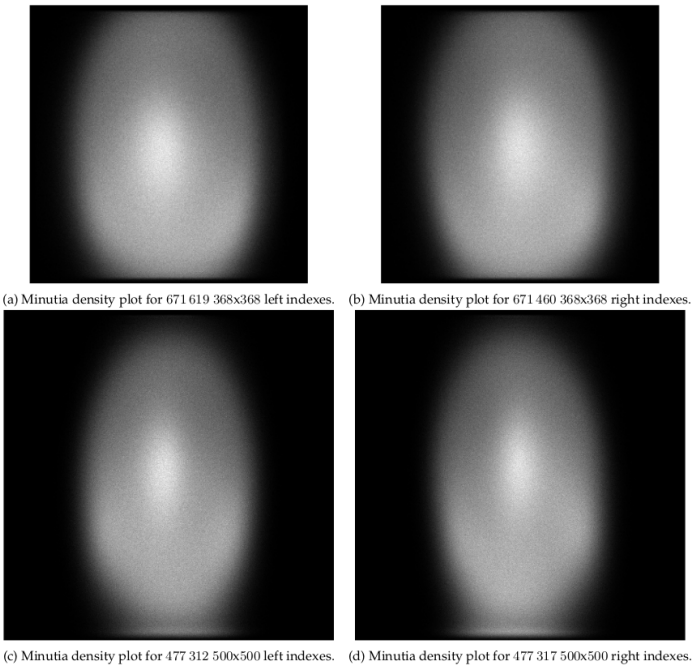
5.2 Results

The results are expressed in terms of Equal-Error Rate (EER) and the periodicity value (Z) computed from the minutiae density plots. The EER at the threshold t is obtained when both False Match Rate and False Non-Match Rate are identical: FMR(t)= FNMR(t). This score was computed, using the FVC Fingerprint Verification Protocol[3] and the matching algorithm from the INCM. Since the Z value is only possible to obtain with a significant amount of data, SFinge were used. In Figure 4 the results obtained in the NIST's MINEX III report are presented. The complete evaluation can be extracted from the MINEX III results page.

[3] https://biolab.csr.unibo.it/FVCOnGoing/UI/Form/BenchmarkAreas/BenchmarkAreaFV.aspx

Dataset	Average EER (baseline)	Average EER (periodicity removal)
FVC	2.528	2.608
800k	1.807	1.926
	Z (baseline)	Z (periodicity removal)
800k	0.0020	0.0010

Table 1: The obtained results before and after the proposed removal periodicity algorithm.



(a) Minutia density plot for 671 619 368x368 left indexes. (b) Minutia density plot for 671 460 368x368 right indexes.

(c) Minutia density plot for 477 312 500x500 left indexes. (d) Minutia density plot for 477 317 500x500 right indexes.

Figure 4: The obtained results of fingerIDAlg after removing the periodicity in the minutiae structure in the NIST's MINEX III.

6 Conclusion

The step for removing periodicity from the detection minutiae phase represents a trade-off between precision and periodicity without influencing the processing time. The experimental testing conducted – see Table 1 and the results in NIST's MINEX III report – see Figure 4 – reveal a significant reduction in the Z value without compromise the accuracy and the performance.

7 Acknowledgements

This work is financed by the ERDF – European Regional Development Fund through the Operational Programme for Competitiveness and Internationalisation - COMPETE 2020 Programme and by National Funds through the Portuguese funding agency, FCT - Fundação para a Ciência e a Tecnologia within project POCI-01-0145-FEDER-030263.

References

[1] R. Cappelli, D. Maio, D. Maltoni, J.L. Wayman, and A.K. Jain. Performance evaluation of fingerprint verification systems. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 3–18, 2006.

[2] R. Cappelli, M. Ferrara, A. Franco, and D. Maltoni. Fingerprint verification competition 2006. In *Biometric Technology Today*, pages 7–9, 2007.

[3] D. Maio, D. Maltoni, R. Cappelli, J.L. Wayman, and A.K. Jain. FVC2000: Fingerprint verification competition. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 402–412, 2000.

[4] D. Maio, D. Maltoni, R. Cappelli, J.L. Wayman, and A.K. Jain. FVC2002: Second fingerprint verification competition. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 811–814, 2002.

[5] D. Maltoni, D. Maio, A.K. Jain, and S. Prabhakar. Handbook of fingerprint recognition. In *Springer Professional Computing*, 2009.

[6] Ana Rebelo, Tiago Oliveira, Manuel E. Correia, and Jaime S. Cardoso. Are deep learning methods ready for prime time in fingerprints minutiae extraction? In *Proceedings of the 23rd Iberoamerican Congress on Pattern Recognition (CIARP)*, 2018. URL publications/conferences/2018ARebeloCIARP.pdf.

[7] E. Tabassi, P. Grother, W. Salamon, and C. Watson. Minutiae interoperability. In *Proceedings of the Special Interest Group on Biometrics and Electronic Signatures*, 2009.

Classifying acanthocytes using image processing and ML techniques: A comparative study

Catarina Silva[1]
c.alexandracorreia@ua.pt
Augusto Silva[1,2]
augusto.silva@ua.pt
Joaquim Madeira[1,2]
jmadeira@ua.pt

[1] Departamento de Electrónica, Telecomunicações e Informática, Universidade de Aveiro
[2] Institute of Electronics and Informatics Engineering of Aveiro, Universidade de Aveiro

Abstract

The diagnosis of several diseases can be improved with the identification of acanthocytes, i.e., red blood cells with abnormal form. We propose an approach to autonomously identify such cells in blood sample images. Our method relies on image processing operations and conventional machine learning methods. The principal motivation is the fact that this identification is usually performed by specialized devices or done manually by humans. Specialized devices are rare and costly, while manual identification is prune to error. Our approach reaches a precision of 91%, showing the potential of the solution.

1 Introduction

Red Blood Cells (RBC) are the most common cells present in the human body [5]. Normal RBC usually have a biconcave disk shape. If there is any abnormality in the shape of RBC, then it may indicate the presence of a disease. Furthermore, the shape and number of anomalous cells may also be an important indicator for medical diagnosis, improving its accuracy. It is important to segment and classify anomalous blood cells in order to detect diseases in a early stage, increasing the chances of successful recovery [1].

There are several types of anomalous blood cells, however we have focused our efforts on acanthocytes. The manual classification of abnormal cells under the microscope tends to give inaccurate results and errors [1]. Autonomous systems to detect and classify abnormal cells reduce the time needed to accomplish such task [3]. Furthermore, the latter typically have a lower error rate when compared to humans for that kind of repetitive work.

Our main objective is to develop a reliable detection and classification procedure for acanthocytes, using a reduced set of features. Image processing techniques are used to segment blood cells and conventional Machine Learning (ML) models to classify them. The output is the classification of each blood cell into one of two classes: normal cells or acanthocytes. Additionally, the number of acanthocytes in the blood sample is computed.

This paper is organised as follows: Section 2 presents the relevant background; in Section 3 the stages of the proposed approach are described; some details regarding the implementation are presented in Section 4; results are presented in Section 5; Section 6 presents some conclusions and ideas for future work.

2 Background

RBC suffer anomalies related with shape, size and color. According to [7], acanthocytes are “Erythrocytes with a dented and prickly profile with spicules of different lengths”. The presence of acanthocytes is a strong indicator of several diseases, such as alcoholic cirrhosis, neonatal hepatitis and poor absorption states.

Several authors have developed methods to autonomously detect the presence of acanthocytes in medical images [4]. Two recent works [7, 8] proposed solutions based on image processing and classification methods.

The first work [7] applies morphological operations to extract the contour of the RBC and computes several features related with contour shape, such as: chain code, circularity and skeleton. After that, they use k-NN [9] as a classification algorithm to classify the extracted contours.

The second work [8] relies on image segmentation as a method to correctly extract the region of each individual blood cell. They also use ML methods for the final classification, namely k-NN and SVM [9].

3 Proposed approach

The proposed approach is based on [7] with some key differences. The image processing workflow is enhanced with the goal of improving region segmentation and contour extraction. We consider a reduced set of features that still contains enough information to properly classify the RBC while speeding-up the ML training time. Finally, we evaluated several ML models instead of only relying on k-NN.

The image processing pipeline is composed by several stages with the objective of reducing noise, enhancing region contours and segmenting them. The key stages of the pipeline are depicted in Figure 1.

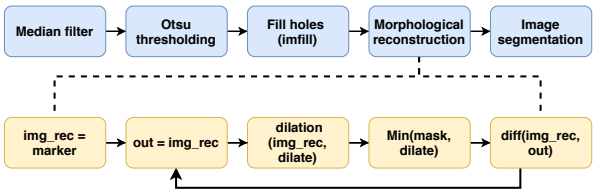
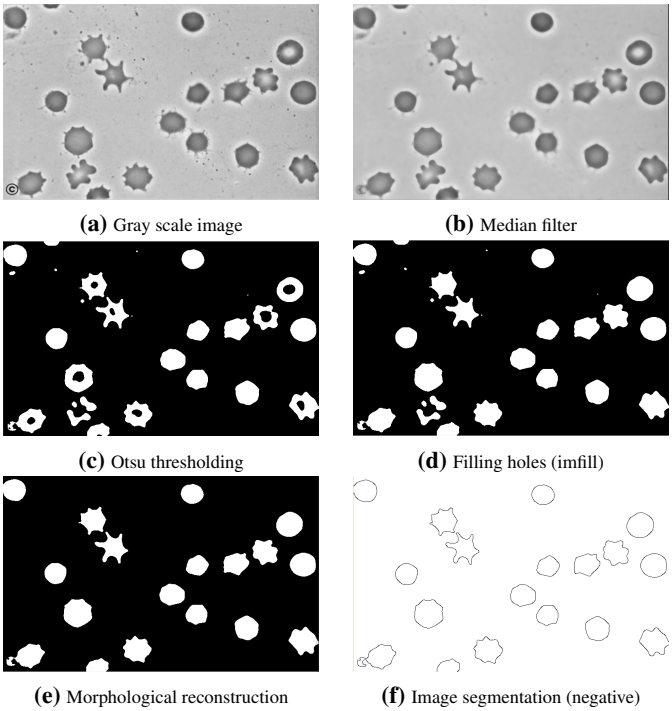


Figure 1: Image processing pipeline.

The first step is to normalize an input image by converting it to gray scale and applying a 9x9 median filter to smooth noise. The gray image is then converted to binary using the Otsu thresholding method. Those operations may originate some holes in the middle of the cells and medium-sized noise (by-product of the binarization).

The next steps fix that by executing a filling operation (imfill) that applies a guided flooding operation to close holes inside blobs. Morphological reconstruction (elliptic shaped 9x9 kernel) is applied to remove the medium-sized noise produced during the binarization. Finally, the Canny edge detector is applied to extract region contours. Figure 2 illustrates the results of the image processing pipeline.



(a) Gray scale image **(b)** Median filter

(c) Otsu thresholding **(d)** Filling holes (imfill)

(e) Morphological reconstruction **(f)** Image segmentation (negative)

Figure 2: Result of the image processing pipeline.

Some care is needed when selecting features: since we are dealing

with images, it is important to compute features that correctly express shape characteristics but are also invariant to rotation and scaling [2].

Based on the extracted region contours, we compute several features that describe them. The first feature is the histogram from the chain code. A chain code characterizes the shape of a contour but is not rotation and scale invariant. To achieve that we compute an histogram with the relative weight for each direction of the chain code. The remaining features are circularity, roundness, aspect-ratio and solidity. The previously mentioned features are shape descriptors commonly used by image processing toolboxes to classify blobs. These features are meant to enhance the classification process, by expanding the expressiveness of the histogram, and capture characteristics that are invariant to scale and rotation.

The chain code histogram, captures the complete shape into a condensed form. The remaining features capture the smoothness and stability of the shape. A smooth round blood cell will have a high circularity value, an aspect-ratio close to one and a high solidity value. The irregular shape of an acanthocyte will diverge from the previously mentioned values.

4 Implementation

The project code was developed in C++ using the Open Source Computer Vision (OpenCV) library[1] and it is hosted on the GitHub public repository[2].

All the image processing and feature extraction code was fully developed. It is important to mention that the imfill[3] and morphological reconstruction operations do not exist in OpenCV and were implemented. Furthermore, OpenCV offers limited support regarding chain codes, we implemented a method to follow each of the contour pixels and build a chain code sequence. Two classifiers were also implemented: k-NN and Logistic Regression.

The current prototype is composed by two main programs: one uses a dataset to train the previously mentioned models, the second uses the model to classify and count acanthocytes in blood cell images. In order to evaluate our prototype using other ML methods we implemented a method that outputs the features into a ARFF[4] dataset. An ARFF dataset can then be loaded into WEKA [6], a popular ML framework.

5 Results

We built a dataset for the evaluation of our prototype, by gathering medical images from microscopic blood samples. Based on the definition of acanthocytes, and on input from medical professionals, we manually segmented those images and classified each blood cell into the healthy and acanthocyte classes. The dataset is composed of 140 segmented images, 72 samples for the acanthocyte class and the remaining 68 samples as the healthy blood cells. The segmented images were resized to have a height of 96 pixels while maintaining the aspect ratio. The dataset is publicly available and can be found on the repository alongside the code.

The segmented images are processed by our proposed image processing pipeline, generating a ARFF output. After that we used the WEKA framework to explored and evaluate several ML models (not being limited by the kNN and logistic regression previsouly developed). It is important to mention that the default hyper-parameters were used for each model.

Three different metrics were used to evaluate the performance of the models: Precision, F-Measure and Matthews correlation coefficient (MCC). The models were evaluated with 10-fold cross validation. The results can be found on Table 1

All models achieve close to 70% precision demonstrating the potential of the developed approach. The top three models are: Random Forest, Neural Network (multi-layer perception) and Decision Tree. One interesting aspect regarding the decision tree model is that the model only uses 5 features: solidity, circularity, aspect ratio, h5 and h3 (h0 to h7 are the values from the chain code histogram). In other words, the model selects as relevant features solidity and circularity, while the remaining three (aspect ratio, h5 and h3) are used only for corner cases. Furthermore, it does not rely on the full histogram for the classification and only uses two of the eight available values.

[1] https://opencv.org/

[2] https://github.com/catarinaacsilva/medical-image-processing

[3] https://www.mathworks.com/help/images/ref/imfill.html

[4] https://www.cs.waikato.ac.nz/ ml/weka/arff.html

Table 1: Classifier performance

ML Algorithm	Precision	F-Measure	MCC
k-NN(1)	0.710	0.704	0.415
k-NN(3)	0.709	0.684	0.400
k-NN(5)	0.748	0.723	0.476
Naive Bayes	0.680	0.652	0.342
Logistic Regression	0.867	0.864	0.731
Decision Tree	0.879	0.879	0.757
Random Forest	**0.910**	**0.909**	**0.819**
Support Vector Machine	0.711	0.630	0.363
Neural Network	0.886	0.886	0.773

6 Conclusions

We proposed a new approach for acanthocyte detection and classification based on image processing and ML models. Contrary to the current trend, we achieved a high precision without relying on Deep Neural Networks, that require substantial amount of data and time to train effectively.

The proposed prototype achieves 91% precision, demonstrating the potential of the solution. We intend to improve our prototype by testing other image segmentation methods (*e.g.* watershed) and convert the code into Python to leverage the advanced ML frameworks available. Furthermore, we intend to explore the importance of each feature that composes our dataset and devise new ones that can improve the accuracy.

References

[1] H. A. Aliyu, R. Sudirman, M. A. Abdul Razak, and M. A. Abd Wahab. Red blood cell classification: Deep learning architecture versus support vector machine. In *2nd Int Conf on BioSignal Analysis, Processing and Systems (ICBAPS)*, pages 142–147, 2018.

[2] Mário Antunes and Luís Seabra Lopes. Unsupervised internet-based category learning for object recognition. In *Lecture Notes in Computer Science*, pages 766–773. Springer, 2013.

[3] S. F. Bikhet, A. M. Darwish, H. A. Tolba, and S. I. Shaheen. Segmentation and classification of white blood cells. In *IEEE Int Conf on Acoustics, Speech, and Signal Processing*, volume 4, pages 2259–2261, 2000.

[4] Evangelia Christodoulou, Jie Ma, Gary S. Collins, Ewout W. Steyerberg, Jan Y. Verbakel, and Ben Van Calster. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *Journal of Clinical Epidemiology*, 110:12 – 22, 2019.

[5] P. T. Dalvi and N. Vernekar. Computer aided detection of abnormal red blood cells. In *IEEE Int Conf on Recent Trends in Electronics, Information Communication Technology (RTEICT)*, pages 1741–1746, 2016.

[6] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. The WEKA data mining software. *ACM SIGKDD Explorations Newsletter*, 11(1):10–18, nov 2009.

[7] María Elena Cruz Meza, MengYi En, Graciela Vázquez Álvarez, and José Cruz Martínez Perales. Detection and classification of abnormalities in erythrocytes by techniques of image analysis and pattern recognition. In *16th LACCEI International Multi-Conference for Engineering, Education, and Technology*, pages 18–20, July 2018.

[8] Cecilia Di Ruberto, Andrea Loddo, and Lorenzo Putzu. A region proposal approach for cells detection and counting from microscopic blood images. In *Lecture Notes in Computer Science*, pages 47–58. Springer, 2019.

[9] Grigorios Tsoumakas and Ioannis Katakis. Multi-label classification. *International Journal of Data Warehousing and Mining*, 3(3):1–13, jul 2007.

Lifelog Moment Retrieval Web Application

Ricardo Ribeiro
rfribeiro@ua.pt
Alina Trifan
alina.trifan@ua.pt
José Luis Oliveira
jlo@ua.pt
António J. R. Neves
an@ua.pt

IEETA/DETI
University of Aveiro
3810-193 Aveiro, Portugal

Abstract

This paper presents an approach for a lifelog moment retrieval application able to provide a visual and interactive environment to the user. This application is divided into three main modules: 1) the user uploads images, as well as eventual textual data annotations; 2) the user interacts with the application introducing relevant words to retrieve a specific moment and, consequently, the application retrieves the images associated to the moment with a certain confidence; 3) a visual environment is provided with two different views, in the form of an image gallery or data tables organized into timestamp clusters. Experimental results confirms the ability to retrieve images from desired moments in personal lifelogs and was used with success in the ImageCLE Flifelog challenge.

1 Introduction

Lifelogging is the process of tracking and recording personal data created through our activities and behaviour [1], consisting of a unified digital record of the totality of an individual's experiences, captured multimodally through digital sensors and stored permanently as a personal multimedia archive. Personal lifelogs have a great potential in numerous applications, including memory and moments retrieval, daily living understanding, or disease diagnosis, as well as other emerging application areas [2]. For example: in Alzheimer's disease, people with memory problems can use a lifelog application to help a specialist follow the progress of the disease, or to remember certain moments from the last days or months.

One of the biggest challenges of lifelog applications is the large amount of data that a person can generate. The lifelog datasets, for example the ImageCLEFlifelog dataset [3], are rich multimodal datasets which consist in one or more months of data from multiple lifeloggers. Therefore, an important aspect is the lifelog data organization in the interest of improving the search and retrieval of information. In order to organize the lifelog data, useful information has to be extracted from it. Other important aspects are the visualization and user interface of the application.

This paper proposes a new concept for a lifelog web application designed and developed for the participation in the ImageCLEF lifelog task [3], more specifically in the Lifelog Moment Retrieval (LMRT) sub-task. This web application was developed in order to visualize and provide an interactive tool to the users. The application is divided into three blocks, namely upload, retrieval and visualization. Each block provides interaction with the user.

2 Related Work

Over the last few years, the term lifelogging has received significant attention from both research and commercial communities. In order to increase the interest on this topic, several challenges started to emerge providing test collection for personal lifelog data [3, 4, 5]. These challenges promote a comparative evaluation of information access and retrieval systems operating over personal lifelog data.

In a Microsoft research project, a lifelog retrieval engine based on an underlying database system, named MyLifeBits [6] was developed. This is generally considered as the first lifelog retrieval system. Zhou *et al.* [7] proposed an iterative lifelog search engine, called LIFER, that is queried based on several different forms of lifelog data, such as visual concepts, activities, locations, time, etc. A retrieval and exploration tool was presented by Leibetseder *et al.* [8], called lifeXplore, that allows to search and browse features that have been optimized for lifelog data. In addition to these works, several others works were presented in lifelog challenges, such as Exquisitor [9] and LifeSeeker [10].

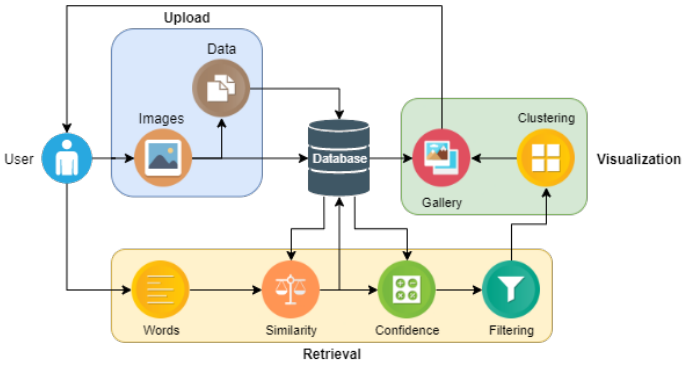


Figure 1: General representation of the developed application. The user interacts with the three blocks: Upload, Retrieval and Visualization.

3 Web Application

The first version of a web application was developed in order to visualize and provide an interactive tool for users. As a proof of concept, the data provided by the ImageCLEFlifelog organizers [3] was used in the application. This web application was divided into three mains modules, such as upload, retrieval and visualization blocks, which provide interaction with the user in each one. In Figure 1 is presented a general representation of the web application.

3.1 Upload

The user uploads the images into the application. Initially in order to test and save time on this module, the textual data annotations provided by the ImageCLEFlifelog organizers are automatically uploaded and organized in the application database associated with uploaded images. However, our application is able to extract annotations from the input images using several methods for image classification, object detection and scene recognition.

The data is organized in the application database into different tables/models, such as images, concepts, locations, activities, scenes, attributes, among others. Each model maps to a single database table. The relationship between models makes our system faster and more efficient. The image model has a many-to-many relationship with the other models. For example: an image can contain several concepts, and a concept can be found in several images.

3.2 Retrieval

This approach only computes the confidence of some images that are selected in a first step for the specific moment by using the cosine similarity of word vectors, which makes this retrieval method more efficient and using less processing time compared with a exhaustive method [11].

In this application, the user retrieves a specific moment from his personal lifelogs by introducing relevant words divided into several categories, such as objects, locations, activities and irrelevant words. If the desired moment contains time ranges, years or days of the week, the user

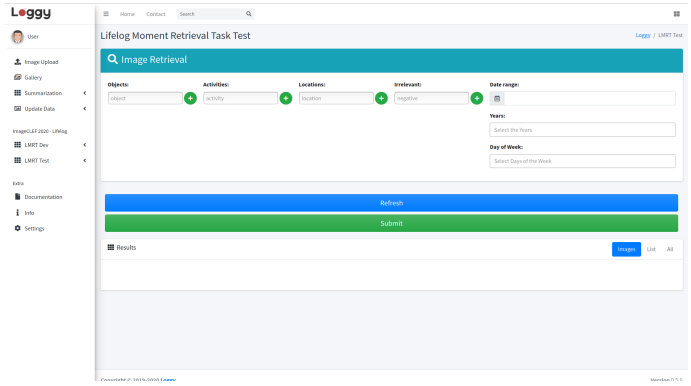


Figure 2: Web application retrieval view.

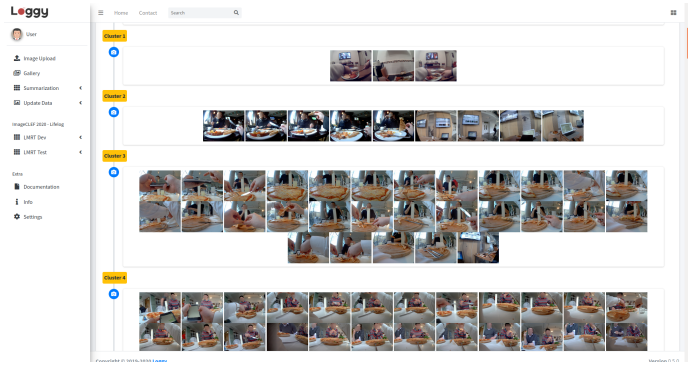


Figure 3: User view of the image clusters in form of image galleries.

can also insert that data in our application to further filter the retrieved images. Figure 2 shows the retrieval view of the web application. In the retrieval block, the input arguments are: objects that appear on the images; activities that the user was practicing; locations or places where the user was; negatives or irrelevant things, activities or locations that should not appear in the images; time ranges, years and days of the week (Monday, Tuesday, Wednesday, Thursday, Friday, Saturday, and Sunday).

The confidence of the selected images is computed using the cosine similarity of word vectors and the score of the labels. For labels without score field, it is only used the similarity to calculate the confidence. As last filtering on the retrieval block, the images are selected based on the confidence threshold, which can be adjuster by the user.

3.3 Visualization

The selected images are organized into different clusters based on images timestamps. The application provides an easy way for users to visualize and identify the clusters that are associated to the specific topic. Figure 3 shows the user view of the clustered images in form of images gallery. Another way of visualization in form of a data table is provided as shown in Figure 4.

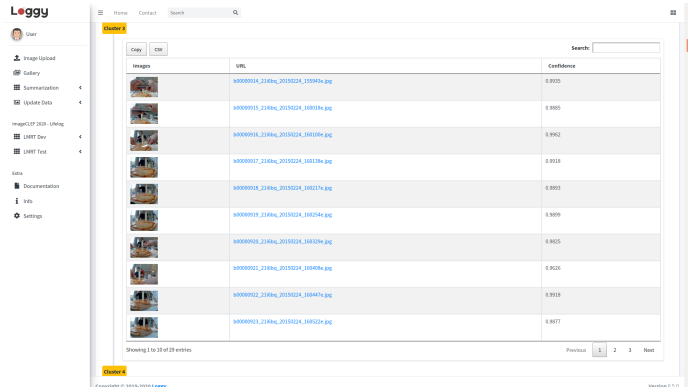


Figure 4: User view of the image clusters in form of data tables.

4 Conclusion and Future Work

This paper presents a proposed concept for a lifelog application, this approach that aims to help people to improve their quality of life. Using only the ImageCLEFlifelog dataset [3] leaves this application somewhat limited regarding the visual concepts extracted from the images. However, using the most recent state-of-art algorithms, a more rich description of the images can be obtained, resulting in an increase of performance. Therefore, it is pretended to implement better scene recognition, object detection, activity and color detection algorithms for the new version of the application.

In future versions, pre-processing methods will be implemented in the application, such as selecting images in upload stage based on low level properties [12]. Moreover, using other metadata acquired together with the images, such as GPS coordinates, it is also possible to improve the performance of the application.

References

[1] Martin Dodge and Rob Kitchin. 'outlines of a world coming into existence': pervasive computing and the ethics of forgetting. *Environment and planning B: planning and design*, 34(3):431–445, 2007.

[2] Peng Wang, Lifeng Sun, Alan F Smeaton, Cathal Gurrin, and Shiqiang Yang. Computer vision for lifelogging: Characterizing everyday activities based on visual semantics. In *Computer Vision for Assistive Healthcare*, pages 249–282. Elsevier, 2018.

[3] Van-Tu Ninh et al. Overview of ImageCLEF Lifelog 2020: Lifelog Moment Retrieval and Sport Performance Lifelog. In *CLEF2020 Working Notes*, CEUR Workshop Proceedings, Thessaloniki, Greece, September 22-25 2020.

[4] Cathal Gurrin, Xavier Giro-i Nieto, Petia Radeva, Mariella Dimiccoli, Duc-Tien Dang-Nguyen, and Hideo Joho. Lta 2017: The second workshop on lifelogging tools and applications. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1967–1968, 2017.

[5] Cathal Gurrin, H. Joho, Frank Hopfgartner, Liting Zhou, Tu Ninh, Tu-Khiem Le, Rami Albatal, D.-T Dang-Nguyen, and Graham Healy. Overview of the ntcir-14 lifelog-3 task. In *Proceedings of the Fourteenth NTCIR conference*, 06 2019.

[6] Jim Gemmell, Gordon Bell, and Roger Lueder. Mylifebits: a personal database for everything. *Communications of the ACM*, 49(1):88–95, 2006.

[7] Liting Zhou, Zaher Hinbarji, Duc-Tien Dang-Nguyen, and Cathal Gurrin. Lifer: An interactive lifelog retrieval system. In *Proceedings of the 2018 ACM Workshop on The Lifelog Search Challenge*, pages 9–14, 2018.

[8] Andreas Leibetseder, Bernd Münzer, Manfred Jürgen Primus, Sabrina Kletz, Klaus Schoeffmann, Fabian Berns, and Christian Beecks. lifexplore at the lifelog search challenge 2019. In *Proceedings of the ACM Workshop on Lifelog Search Challenge*, pages 13–17, 2019.

[9] Omar Shahbaz Khan, Bjorn Thor Jonsson, Jan Zahalka, Stevan Rudinac, and Marcel Worring. Exquisitor at the lifelog search challenge 2019. In *Proceedings of the ACM Workshop on Lifelog Search Challenge*, pages 7–11, 2019.

[10] Tu-Khiem Le, Van-Tu Ninh, Duc-Tien Dang-Nguyen, Minh-Triet Tran, Liting Zhou, Pablo Redondo, Sinead Smyth, and Cathal Gurrin. Lifeseeker: Interactive lifelog search engine at lsc 2019. In *Proceedings of the ACM Workshop on Lifelog Search Challenge*, pages 37–40, 2019.

[11] Ricardo Ribeiro, António JR Neves, and José Luis Oliveira. Ua.pt bioinformatics at imageclef 2019: Lifelog moment retrieval based on image annotation and natural language processing. In *CLEF (Working Notes)*, 2019.

[12] Ricardo F Ribeiro, António JR Neves, and José Luis Oliveira. Image selection based on low level properties for lifelog moment retrieval. In *Twelfth International Conference on Machine Vision (ICMV 2019)*, volume 11433. International Society for Optics and Photonics, 2020.

Achieving Cancellability in End-to-End Deep Biometrics with the Secure Triplet Loss

João Ribeiro Pinto [1,2]
joao.t.pinto@inesctec.pt
Miguel V. Correia [1,2]
mcorreia@fe.up.pt
Jaime S. Cardoso [1,2]
jaime.cardoso@inesctec.pt

[1] INESC TEC
Porto, Portugal
[2] Faculdade de Engenharia
Universidade do Porto
Porto, Portugal

Abstract

The literature on biometric recognition shows a chasm between the methods focused on high performance and the works focused on template security. To build a connection between these two worlds, this work describes the Secure Triplet Loss to achieve template cancellability within end-to-end deep learning models. Evaluated for off-the-person electrocardiogram authentication, the proposed methodology resulted in effective cancellability, irreversibility, and improved performance. Despite the high linkability, this shows that it is possible to combine the high performance of deep learning with adequate template security.

1 Introduction

As biometric recognition technologies quickly conquer a place of relevance in our society, the duality of performance *versus* security is yet to be adequately addressed [11]. This duality relates to how the research field of biometrics is currently composed of two 'worlds' apart, and while both work towards the same goal of improving human recognition systems, they have been following largely unconnected and uncoordinated research lines.

On the one hand, a substantial part of the literature in biometrics is focused on performance, following well-known and successful methodologies in computer vision tasks. These use mostly end-to-end convolutional neural networks (CNNs) that consider biometric recognition as a general classification problem [14], and have achieved outstanding levels of accuracy and robustness in challenging scenarios. However, since stored data protection is rarely addressed, these algorithms are incomplete and unfit for real biometric applications.

On the other hand, several algorithms have been proposed to protect personal data stored in biometric systems [5, 8]. These commonly use cryptography and information theory concepts to ensure stored biometric templates verify the essential properties of irreversibility, cancellability, and non-linkability. Nevertheless, being based on separate processes means these methodologies are not applicable to state-of-the-art end-to-end methods without significant negative impacts on performance. This is a relevant problem, since many biometric traits (including the electrocardiogram, ECG) rely on end-to-end CNNs to offer acceptable accuracy and robustness to challenging scenarios [9].

This work aims to bring the two aforementioned research lines together by answering the following question: *if deep learning models have successfully learnt so many different things, why not template security?* The proposed method is an adaptation of the triplet loss [2], which aims to achieve template irreversibility and cancellability on end-to-end CNNs while preserving recognition accuracy. This methodology is used to train a competitive end-to-end model for ECG biometric recognition [9] and evaluated on the off-the-person UofTDB database [13]. Thus, this work addresses the challenge of template protection on end-to-end networks for ECG and biometrics in general, contributing towards a synergy between performance and security in biometric recognition.

2 The Secure Triplet Loss

The triplet loss [2] is used to train models to determine whether or not two samples belong to the same class [3, 4, 9]. The model receives a triplet of inputs: an anchor (x_A of class i_A), a positive sample (x_P of class $i_P = i_A$), and a negative sample (x_N of class $i_N \neq i_A$). Considering the case of biometric recognition, the samples are biometric trait measurements and the classes are identities.

The model will output an embedding y for each input (*e.g.*, $y_A = f(x_A)$ for the anchor). Two embeddings can be compared through a metric of distance or dissimilarity $d(y_1, y_2)$ which can be used to determine if the respective inputs belong to the same class. The model can be trained through the triplet loss

$$l = \max\left(0, \alpha + d(y_A, y_P) - d(y_A, y_N)\right), \tag{1}$$

which will promote the maximisation of $d(y_A, y_N)$ and the minimisation of $d(y_A, y_P)$, grouping samples of the same class into compact clusters, at least α from other classes in the embedding space.

Although the triplet loss has been successfully applied to several pattern recognition problems, including biometric authentication, it does not address the important issue of template cancellability. Typically, this is performed separately, binding a subject-specific key k with the template after it is generated: changing k invalidates any compromised templates bound with other keys.

Here, we adapt the triplet loss to perform subject key binding with the template within the end-to-end model. Besides the biometric samples x_A, x_P, and x_N, the model will receive two keys, k_1 and k_2. Sample x_A is bound with k_1 and x_P and x_N are bound with each of the two keys, resulting in five embeddings: $y_A = f(x_A, k_1)$, $y_{P1} = f(x_P, k_1)$, $y_{P2} = f(x_P, k_2)$, $y_{N1} = f(x_N, k_1)$, $y_{N2} = f(x_N, k_2)$. From these, four distances are computed: $d_{SP} = d(y_A, y_{P1})$ (with matching identities and keys), $d_{DP} = d(y_A, y_{P2})$ (with matching identities but different keys), $d_{SN} = d(y_A, y_{N1})$ (with different identities but matching keys), and $d_{DN} = d(y_A, y_{N2})$ (with non-matching identities and keys).

Since d_{SP}, which corresponds to matching identities and keys, should be minimised, while the others should be maximised, the Secure Triplet Loss is computed through:

$$l = \max\left(0, \alpha + d_{SP} - \min(\{d_{SN}, d_{DP}, d_{DN}\})\right). \tag{2}$$

As with the triplet loss, α will enforce a margin between positive and negative distances. By minimising the loss in Eq. (2), the model learns to deal with the intrasubject and intersubject variability of the biometric trait and becomes able to recognise when the keys do not match, even if the identity is the same. Hence, if the stored templates become compromised, they can easily be invalidated through a key change.

3 Experimental Settings

The proposed training methodology was evaluated to off-the-person ECG-based biometric authentication. The University of Toronto ECG Database (UofTDB) [13], including 1019 identities, was used. Signals were divided into five-second segments. Data from the last 100 subjects were used for training (90 000 triplets) and validation (10 000 triplets), while the data from the remaining 918 subjects were reserved for testing (10 000 triplets). Keys were randomly generated as unidimensional arrays of 100 binary values.

The authentication model is adapted from [9] (see Fig. 1). Samples are bound with keys before the first dense layer. The vector of flattened feature maps ($s(x)$) is concatenated with a key k (after its normalisation to unit $l2$ norm). The last dense layer outputs the respective representation $y = f(s(x), k)$, which is then used in dissimilarity score computation using the Euclidean and normalised Euclidean distance, respectively, for training and testing. The model was trained using the Adam optimizer with an initial learning rate of 0.0001, for a maximum of 500 epochs, with early stopping based on validation loss (patience of 20 epochs).

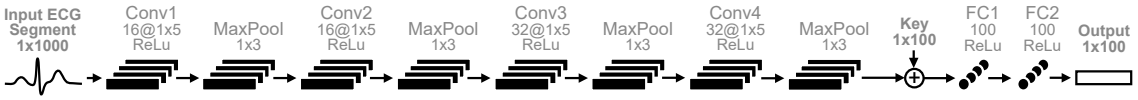


Figure 1: Architecture of the model trained for ECG-based authentication.

4 Results and Discussion

After training, the model's authentication performance was evaluated through the analysis of false acceptance (FAR), rejection rates (FRR), and equal error rates (EER) [10]. As presented in Fig. 2, the model trained with the Secure Triplet Loss achieved lower EER than with the original triplet loss (10.63% versus 12.55%). This is an important aspect of the proposed method, since security measures generally lead to a five-fold average increase in authentication error [8]. The proposed method is able to achieve this by retaining the capabilities of end-to-end networks and optimising for accuracy and cancellability simultaneously.

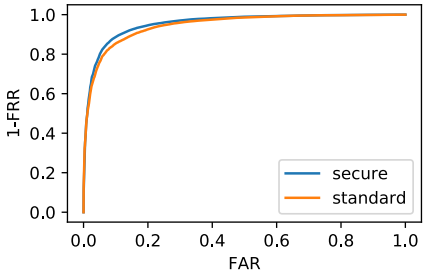


Figure 2: Receiver-operating characteristic curves of the model trained with the Secure Triplet Loss and with the original triplet loss.

The security of the templates output by the model was evaluated using the standard literature measures of privacy leakage rate, secrecy leakage, and secret key rate, through nearest-neighbour entropy estimation methods [1, 6, 7]. The model offered near-perfect privacy rate results, which means the biometric templates are irreversible as desired. This very useful property may be a consequence of using CNNs, which have been observed to present minimal mutual information between inputs and outputs when appropriately optimised [12]. Secrecy leakage also rendered perfect result (0) which may also be related to the nature of deep neural networks. At last, the proposed method offered 103.73 bits of secret key rate (output entropy) versus 14.20 bits for the original triplet loss, which means it will be harder to successfully attack the model trained with the proposed method.

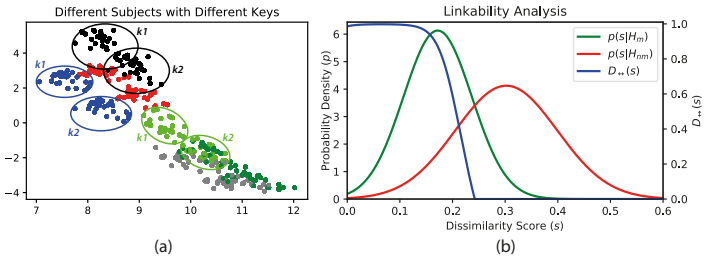


Figure 3: Results of the cancellability (a) and linkability (b) evaluation.

Regarding template cancellability, Singular Value Decomposition (SVD) was used to visualise the template distribution in the output space. Fig. 3 (a) shows the Secure Triplet Loss promotes the clustering of class samples when keys match. However, when the key is changed, the cluster is shifted on the output space in order to distance itself from (and effectively invalidate) the templates corresponding to cancelled keys. At last, template non-linkability was evaluated as established by Gomez-Barrero *et al.* [5] (see Fig. 3 (b)). The proposed secure triplet loss model offered $D_{\leftrightarrow}^{sys} = 0.67$, making it semi- to fully linkable. This is the main shortcoming of the Secure Triplet Loss, as it would be relatively easy for an attacker to discover whether two samples with different keys belong to the same subject. The desired behaviour would be for d_{DP}, d_{DN}, and d_{SN} to assume similar values greater than d_{SP}. Future research endeavours should focus on adapting the network to avoid template linkability.

5 Conclusion

This work proposes the Secure Triplet Loss, an adaptation of the triplet loss to promote biometric template cancellability in end-to-end deep models. Biometric templates are bound with subject-specific keys within the end-to-end model, without separate processes, and can be easily cancelled through a key change. The proposed loss proved successful when evaluated for ECG-based authentication, offering cancellability and improved performance.

While cancellability and irreversibility have been achieved, an important shortcoming regarding template linkability has been unveiled. Hence, further efforts should be devoted to achieve non-linkability alongside cancellability. Nevertheless, this study has shown it is possible to achieve template security within end-to-end deep biometric models, paving the path to a synergy between performance and security in biometrics.

Acknowledgements

This work was financed by the ERDF – European Regional Development Fund through the Operational Programme for Competitiveness and Internationalization - COMPETE 2020 Programme and by National Funds through the Portuguese funding agency, FCT – Fundação para a Ciência e a Tecnologia within project "POCI-01-0145-FEDER-030707", and within the PhD grant "SFRH/BD/137720/2018". The authors wish to thank the administrators of the UofTDB database used in this work.

References

[1] P. Brodersen. Entropy estimators, 2017. URL `https://github.com/paulbrodersen/entropy_estimators`.

[2] G. Chechik, V. Sharma, U. Shalit, and S. Bengio. Large scale online learning of image similarity through ranking. *JMLR*, 11, 2010.

[3] W. Chen, X. Chen, J. Zhang, and K. Huang. Beyond triplet loss: A deep quadruplet network for person re-identification. In *CVPR*, 2017.

[4] D. Cheng, Y. Gong, S. Zhou, J. Wang, and N. Zheng. Person Re-Identification by Multi-Channel Parts-Based CNN With Improved Triplet Loss Function. In *CVPR*, 2016.

[5] M. Gomez-Barrero, C. Rathgeb, J. Galbally, C. Busch, and J. Fierrez. Unlinkable and irreversible biometric template protection based on bloom filters. *Information Sciences*, 370-371:18–32, 2016.

[6] L. F. Kozachenko and N. N. Leonenko. Sample estimate of the entropy of a random vector. *Problemy Peredachi Informatsii*, 23, 1987.

[7] A. Kraskov, H. Stögbauer, and P. Grassberger. Estimating mutual information. *Phys. Rev. E*, 69:066138, Jun 2004.

[8] K. Nandakumar and A. K. Jain. Biometric Template Protection: Bridging the performance gap between theory and practice. *IEEE Signal Processing Magazine*, 32(5):88–100, 2015.

[9] J. R. Pinto and J. S. Cardoso. A end-to-end convolutional neural network for ECG-based biometric authentication. In *BTAS*, 2019.

[10] J. R. Pinto, J. S. Cardoso, and A. Lourenço. Evolution, Current Challenges, and Future Possibilities in ECG Biometrics. *IEEE Access*, 6:34746–34776, 2018.

[11] J. R. Pinto, J. S. Cardoso, and M. V. Correia. Secure Triplet Loss for End-to-End Deep Biometrics. In *IWBF*, April 2020.

[12] N. Tishby and N. Zaslavsky. Deep learning and the information bottleneck principle. In *ITW*, April 2015.

[13] S. Wahabi, S. Pouryayevali, S. Hari, and D. Hatzinakos. On Evaluating ECG Biometric Systems: Session-Dependence and Body Posture. *IEEE TIFS*, 9(11):2002–2013, Nov. 2014.

[14] M. Wang and W. Deng. Deep face recognition: A survey. *arXiv*, 2018. 1804.06655.

Image Quality Assessment of Cytology Images using Deep Learning

Tomé Albuquerque[1,2]
tome.m.albuquerque@inesctec.pt

Maria João M. Vasconcelos[3]
maria.vasconcelos@fraunhofer.pt

Jaime S. Cardoso[1,2]
jaime.cardoso@inesctec.pt

[1] Faculty of Engineering of the University of Porto
Porto, Portugal

[2] INESC TEC
Porto, Portugal

[3] Fraunhofer Portugal AICOS
Porto, Portugal

Abstract

The massive growth of digital color image contents in medical field, due to the spread of advanced multimedia devices capable of acquisition, transmission, and storage of digital data, demands improvements in image quality assessment (IQA) methods. Cervical cancer ranks as the fourth most common cancer among females worldwide with roughly 528,000 new cases yearly. Significant progress in the realm of artificial intelligence, particularly in neural networks and deep learning, helps physicians to classify cervical cancer more accurately using cytology and colposcopy images. However, it is necessary to ensure good image quality for decent performance of classifying methods. In this paper, we address a binary IQA problem (bad versus good quality) of cytology images with three different widely used architectures: VGG16, MobileNet, and ResNet50. The experimental results show the good performance of deep learning algorithms for IQA.

1 Introduction

Medical image quality assessment plays an important role not only in the design and manufacturing processes of image acquisition but also in the optimization of decision systems. Cervical cancer remains the fourth most common cause of cancer death in women worldwide. Despite the outburst of recent scientific advances to find an effective treatment, there is no effective method, especially when diagnosed in an advanced stage. However, screening tests such as cytology or colposcopy, have been responsible for a strong decrease in cervical cancer deaths.

Cytology microscope images need high-level microscopic magnification for a consistent characterization, but it is necessary to preserve an appropriate image quality [1]. In most of the cases, the cells in test slides are frequently spread in a multi-layer way which raises a challenge for a good focusing. Thus, it is necessary to use different focus levels for correct digital representation with good image quality. Powerful auto-focusing techniques using IQA methods are used in automated microscopy to prevent the loss of image quality [2]. The adequacy assessment of the cytology image has been studied by several researchers which propose different approaches to guarantee the The Bethesda System (TBS) minimum criteria (cellularity, obscuring factors, and the evidence of transformation zone) [3]. Most of the works on literature discard the importance of IQA and are focused on classification problems in cytology. Therefore, due to the lack of scientific work regarding the IQA on cervical cancer screening method, it is necessary to fulfil that gap with more research on this field to improve the actual state-of-the-art.

1.1 Brief summary

In this work, some approaches for non-reference image quality assessment (NR-IQA) are presented using feature extraction and learning. Thus, different convolutional neural network (CNN) based models, pre-trained on ImageNet dataset, were used and fine-tuned to predict the quality score value of several images. The first models uses three different architectures VGG16, MobileNet, and ResNet50) to predict the image quality of a blood cells microscopy dataset, labelled in a multiclass problem (4 classes).

Due to the lack of annotated cytology dataset regarding image quality, it was used as reference a microscopic blood cell sample dataset. After selecting the best model, the weights of the best model were used to initiate the train of a new model to classify IQA on a new IQA dataset created in this work with cytology microscope images. This new dataset contains reference images and distorted images obtained from the reference images. The classification of IQA on the cytology dataset is done on a binary problem (bad quality vs good quality), this classifier intends to learn low-notion quality features by distinguish between original/ reference images and distorted images.

2 Methods

2.1 Dataset

In this work, two different datasets were used in the train of IQA models. The microscope slide preparations of the first dataset, from blood samples, were obtained in Centro Hospitalar de São João (Porto), and digitalized by Fraunhofer AICOS (FhP-AICOS) within the scope of MpDS project [1] using μsmartscope. The blood cell data was annotated by doctors and contain a total of 1854 images divided into 4 different classes according to quality score of the image, which can be bad, fair, good, or excellent quality. The second dataset was collected in Hospital Professor Doutor Fernando Fonseca (Lisbon) within the scope of CLARE project using an updated version of μsmartscope, adapted to acquire samples with 400 times magnification the images were also acquired as the first dataset by FhP-AICOS using μsmartscope. This second dataset consists of 4088 images of microscope pap smear slide preparations (liquid-based citology samples), in which 817 are reference images and 3271 are images generated from the reference images with four different types of distortions (Gaussian noise, blur, salt and pepper noise and speckle noise). Thus, for every image of reference four new images were created with the distortions mentioned before as showed in Figure 1. This new cytology dataset is divided in 2 different classes, distorted images (bad quality images), and reference images (good quality images).

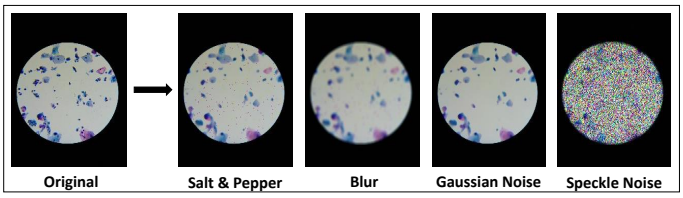


Figure 1: Example of the four different types distortions applied on the original images of cytology dataset.

2.2 Data Pre-processing

First, is worth mentioning that both datasets were divided into train, validation, and test subsets (60/20/20 %) in 5 different k-folds for cross-validation, maintaining the ratio among the different classes. The acquisition of the images in both datasets was done by different experts. In all the images the ROI is a circular region but, in some cases, that region is not in the centre of the image. To guarantee that the input image for the CNN models is a perfect square of the ROI it was necessary do an automatic crop at the limits of the circle creating a square shape image. The images were also resized to 224 x 224 and normalized from the pixel values range of 0-255 to the range 0-1 to speed up the train. Both datasets used in this work had a small number of labelled images (1854/4088), to overcome this obstacle it was used data augmentation. Therefore, during the training of the models a series of random transformations have been done in each training epoch for every single image. The transformations

[1] *https://www.aicos.fraunhofer.pt/en/our_work/portfolio/micron.html*

applied included image rotation with a range of 90°, width and height shift, horizontal flip, zoom (in or out).

2.3 Convolutional Neural Networks

Convolutional neural network (CNN) is a class of deep learning algorithms that are organized in several connected successive stages, specifically convolutional (conv), pooling (pool) layers, and activation functions. CNN has learnable parameters named weights that can be updated using several matrix multiplications and its goal is to reduce the images into a form that is easier to process and classify. After the convolutional block of the neural network, it follows a Fully Connected (FC) network that has as input the flatten feature map with origin in the convolutional output. The last layer of the FC computes the classification probability for each class using SoftMax regression for multiclass classification or sigmoid for binary classification.

3 Experimental Details

Two different models were trained and tested in this work. The first model 1 was tested with three different CNN architectures. All the architectures were directly fed with the blood cells and cytology images. The CNN architecture that achieves the best results in the IQA of blood cells dataset was used on model 2 for cytology IQA. The tested CNN architectures are the following: VGG16, MobileNet and ResNet50. For multiclass IQA of the blood cells, it was implemented the model 1 with the 3 different convolutional architectures mentioned above. Model 2 follows the same pattern of model 1, however, it is a CNN model for binary classification. This model classifies the image quality of cytology cells in two different classes. The pre-trained weights of model 1 are used to initiate the train of model 2 through transfer learning to increase the performance and diminish training time. The hyperparameters were adjusted just for model 1 with the grid search where the hyperparameters tuned were the learning rate (LR $\in$ [0.001; 0.0001]), the batch size (BS $\in$ [16; 32]), and the dropout rate (Dt $\in$ [0.2; 0.5]). During 500 epochs (using ModelCheckPoint and EarlyStopping TensorFlow callbacks) the model 1 was tested with all these hyperparameters. The losses used in this work were categorical and binary cross-entropy for multiclass and binary classification respectively.

4 Results and Discussion

The best combination of the hyperparameters found for model 1 after a fine-tuning was LR of 0.0001, BS of 32, and Dt of 0.2. These hyperparameters were chosen concerning the validation subset. The best results after the grid search for the first model (model 1 - multiclass) are represented in Table 1. For the multiclass problem, VGG16 achieved the best results with higher values in all the presented metrics when compared with MobileNet and ResNet50.

Table 1: Results of model 1 for multiclass classification task of IQA in blood cells test subset.

	Accuracy (%)	Precision (%)	recall (%)	AUC (%)
MobileNet	76.05± 1.74	76.34 ± 2.75	76.05 ± 1.74	86.85 ± 1.39
VGG16	**78.91 ± 1.97**	**79.16 ± 2.17**	**78.91 ± 1.97**	**87.64 ± 1.82**
ResNet50	76.97 ± 2.39	77.29 ± 2.67	76.97 ± 2.39	83.06 ± 3.38

To confirm if the resize of the images to 224 x 224 did not contribute to loss of information an additional model trained with resized images but with bigger dimensions (512 x 512) was created and tested. Due to the imbalance in the number of images per class in blood cells dataset classes a new model was done by oversampling the minority classes. The results for these different approaches using different shows no upgrades in the metrics. The model 1 trained with VGG16 as the convolutional block and using as input 224 x 244 resized images achieved the best performance. Thus, the pre-trained weights of this model 1 will be used by model 2 which will be only trained with a VGG16 architecture for binary IQA of cytology images. The last layer was discarded. The results of model 2 are presented in the following table 2.

Table 2: Results of model 2 using VGG16 architecture to assess image quality of pap smear cells (cytology dataset).

	Accuracy (%)	Precision (%)	recall (%)	AUC (%)
VGG16	99.49 ± 0.72	99.50 ± 0.70	99.49 ± 0.72	99.96 ± 0.07

VGG16 in the multiclass quality assessment of blood cells images achieved the best performance in all metrics. This may have happened due to the relatively small number of images in our dataset used in this work. Thus, the high complexity of that the deeper networks (ResNet50 and MobileNet), may lead to overfit on train. The AUC metric is higher than accuracy, which may indicate that our classifier achieves good performance on the positive class (high AUC) at the cost of a high false negatives rate. The model proposed for classification of the pap smear images quality achieved a good performance for the binary image quality classification task with 99.49% of accuracy. This algorithm classifies the images according to the presence or absence of distortions, this way the classifier is focused on low-level notions of quality. The difference between the metrics of models 1 and 2 can be explained by the fact that in model 1 there are much more natural distortions in the images which increases the classification challenge, while our generated dataset for model 2 only contains 4 different types of distortions. To classify the images taking in account semantic complex concepts it is necessary to provide more information to the model.

5 Conclusions and Future Work

The use of IQA in biomedical applications is essential to help in optimization and improvement not only in image processing techniques but also on diagnostic algorithms. Nevertheless, the use of IQA methods in medical applications is very limited to low notions of quality, such as distortions or noise on the images. It is essential to collect more data and semantic information about image quality to build more robust and accurate algorithms to assess quality. Thus, new studies on cervical cancer using IQA techniques such as deep learning techniques should be encouraged due to the flexibility and capacity of these techniques and due to the literature gap about this subject. In this work it was demonstrated the outstanding performance of deep learning algorithms using CNN for NR-IQA in biomedical databases.

For future works, since a screening system is expected to be able to avoid misclassifying, artifacts must be added to the synthetic dataset of cytology to test the capacity of the CNN classifiers to detect that. In the future, it may be also interesting, add noise only in part of the image, and train with these examples. After that, analyze the activation maps and see if the explanation is consistent with the spatial placement of the noise.

Acknowledgements

This work is financed by the ERDF – European Regional Development Fund through the Operational Programme for Competitiveness and Internationalisation - COMPETE 2020 Programme and by National Funds through the Portuguese funding agency, FCT - Fundação para a Ciência e a Tecnologia within project “POCI-01-0145-FEDER-028857”.

References

[1] Ewert Bengtsson and Patrik Malm. Screening for cervical cancer using automated analysis of pap-smears. *Computational and mathematical methods in medicine*, 2014:842037, 03 2014.

[2] Zhi Lu, Gustavo Carneiro, Andrew Bradley, Daniela Ushizima, Masoud S. Nosrati, Andrea Bianchi, Cláudia Carneiro, and Ghassan Hamarneh. Evaluation of three algorithms for the segmentation of overlapping cervical cells. *IEEE Journal of Biomedical and Health Informatics*, 21:1–1, 01 2016.

[3] Teresa Conceição, Cristiana Braga, Luís Rosado, and Maria Vasconcelos. A review of computational methods for cervical cells segmentation and abnormality classification. *International Journal of Molecular Sciences*, 20:5114, 10 2019.

Explainable Artificial Intelligence for Face Presentation Attack Detection

Wilson Silva[1,2]
wilson.j.silva@inesctec.pt
João Ribeiro Pinto[1,2]
joao.t.pinto@inesctec.pt
Tiago Gonçalves[1,2]
tiago.f.goncalves@inesctec.pt
Ana F. Sequeira[1]
ana.f.sequeira@inesctec.pt
Jaime S. Cardoso[1,2]
jaime.cardoso@inesctec.pt

[1] INESC TEC
Porto, Portugal
[2] Faculty of Engineering, University of Porto
Porto, Portugal

Abstract

The use of deep learning techniques for face presentation attack detection (PAD) is increasingly common due to their ability to reach strong accuracy performances. Nonetheless, the use of complex models such as the ones produced with deep learning techniques raises safety and trust concerns, as one is not able to understand the motifs behind model decisions. Furthermore, traditional metrics of evaluation fall short in terms of capturing the desirable working properties of models, which is particularly worrisome when working in high-regulated areas, such as biometrics. In this work, we propose the use of interpretability techniques to further assess the robustness of face PAD models. Moreover, we also define desirable properties for a face PAD model to have, which can be evaluated through interpretability. Experiments were performed using the ROSE Youtu video collection and showed the additional value of interpretability in the identification of model robustness.

1 Introduction

Nowadays, deep learning algorithms are excelling in most of the artificial intelligence (AI) fields, including in the biometrics and forensics domain. Although these models can indeed achieve incredible performances due to their complexity and flexibility, it is also true that sometimes these performances are obtained by a focus in wrong/biased information instead of domain significant information [4]. Therefore, an evaluation performed based on only traditional metrics may be misleading, making us trust a model that is not robust enough to be deployed in the real-world.

With regards to face PAD, the use of deep learning techniques is also increasingly common [6]. Furthermore, the diversity of presentation attacks that can happen in a real-world scenario increase the importance of checking the robustness of the deep models, as they may focus on attack-specific or spurious information instead of more general features capable of characterising what an attack means [3].

To overcome the limitations of evaluating a face PAD model only with the traditional metrics, we propose in this work the use of interpretability methods to further assess how robust is a model, by checking which information is determining the deep learning model decision. Interpretability or explainability (we use both terms interchangeably) is the process of understanding which features, or which process, led to the machine learning model decision. Doshi-Velez and Kim categorised these techniques into three different groups, namely, pre-, in-, and post-model [2]. In the last years, interpretability research has focused attention on the in- and post-model interpretability groups, i.e., in the proposal of interpretable models by design [9], or in the proposal of interpretability methods to analyse previously built models [1].

In this work, we also assess the fulfilment of important properties defined by Sequeira *et al.* [8], such as (1) explanations for the same sample should be similar whether or not it is seen during training (data swap); and (2) explanations for the same sample should be similar whether or not the model is trained to detect that specific attack (One-Attack vs. Unseen-Attack).

2 Methodology

A presentation attack detection method receives as input a biometric trait measurement and returns as output a prediction of the classification of that measurement as belonging to a living individual (*bona fide*) or as being a spoof attempt to intrude the system (*attack*). In this work, our method consists of an end-to-end convolutional neural network, with its architecture being described in Figure 1. Since the focus of the work is the study on the interpretability of the face PAD model, we chose a relatively simple architecture.

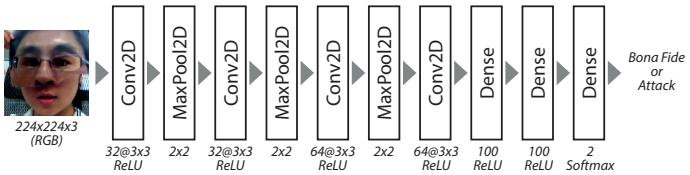


Figure 1: Architecture of the implemented PAD model.

With regards to the interpretability method to be used in this work, we selected the well-known Grad-CAM method [7], as it has the flexibility to generate explanations for any layer of the network, and also allows us to obtain class-specific explanations.

3 Experimental Assessment

The experiments were performed with the ROSE-Youtu Face Liveness Detection Dataset [5], which is composed of 3497 videos acquired from twenty different subjects. For each subject there are several "genuine", and "attack" videos (types of attack, and number of frames extracted are presented in Table 1). The PAD model previously presented was implemented in Keras and trained for 150 epochs with early-stopping (based on validation loss). To avoid overfitting, regularization techniques such as dropout and data augmentation were used.

Table 1: Characteristics of the presentation attack instruments in the ROSE Youtu dataset (N.I. stands for "number of images", *i.e.*, frames extracted from the videos).

Attack	Type of presentation attack instruments	N.I.
-	Genuine (*bona fide*)	2794
#1	Still printed paper	1136
#2	Quivering printed paper	1188
#3	Video of a Lenovo LCD display	923
#4	Video of a Mac LCD display	1113
#5	Paper mask without cropping	1194
#6	Paper mask with two eyes and mouth cropped out	608
#7	Paper mask with the upper part cut in the middle	1162

The quantitative results in terms of *Bona fide Presentation Classification Error Rate (BPCER)*, *Attack Presentation Classification Error Rate (APCER)*, and *Equal Error Rate (EER)* are shown in Table 2. As illustrated in Table 2, we performed the experiments using two different evaluation frameworks: *one-attack* (model is trained and tested with only one type of attack), and *unseen-attack* (model is trained with all but one attack, and tested with the remaining attack). Even though the focus of the work is not on the performance of the face PAD model, the method's performance is in line with state-of-the-art methods. The results with regards to the Unseen-Attack framework are worse than the ones related to the One-Attack framework, which indicate model generalization problems.

Table 2: PAD performance of the models for One-Attack and Unseen-Attack evaluation frameworks. (EER, APCER, and BPCER in %)

Attack	One-Attack			Unseen-Attack		
	EER	*APCER*	*BPCER*	*EER*	*APCER*	*BPCER*
#1	7.29	12.15	3.06	5.90	6.94	4.90
#2	3.62	6.67	1.35	5.55	**3.00**	10.65
#3	2.79	8.37	0.12	10.38	26.29	4.28
#4	12.66	30.38	1.84	25.34	45.73	**3.92**
#5	1.61	**1.61**	1.59	**4.84**	3.55	7.10
#6	4.46	5.10	1.10	10.19	12.74	7.71
#7	**0.73**	5.23	**0.00**	15.49	34.31	7.71

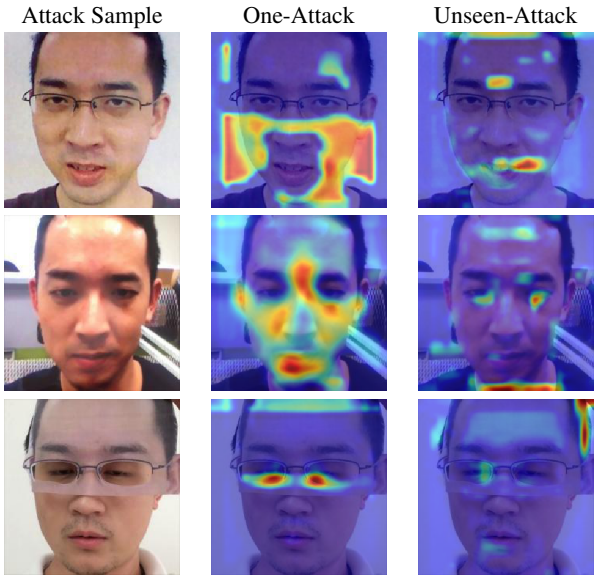


Figure 2: Explanations for correctly classified attack samples (TP) in the One-Attack (2nd column) or Unseen-Attack (3rd column) frameworks. Each row corresponds to one specific type of attack, top to bottom: #1, #4, and #7.

Apart from the usual quantitative evaluation performed for PAD models, we introduce here a qualitative evaluation of model properties based on explanations. With this regard, two types of experiments were performed: comparing explanations for the same attack sample when in the one-attack framework or the unseen-attack framework; and, comparing explanations when attack samples of a random subject are present in train or test (swap experiment). The results obtained with these two approaches are presented in Fig. 2 and Fig. 3. As it can be observed in Fig. 2, the explanations generated for the same samples in the *one-attack* and *unseen-attack* frameworks are quite different, not showing coherence on the information that is relevant to making the decision, which again indicates there are generalization issues with the models. On the other hand, the models demonstrated to be robust with regards to unseen subjects, as the explanations generated in the swap experiment show relevance of the same regions independently of the subject under analysis being in train or test.

4 Conclusions and Future Work

In this work, interpretability techniques were explored to further assess the robustness of face PAD models. Moreover, we studied several desirable properties for a face PAD model to fulfil that are only verifiable through an interpretability analysis of the models. Nonetheless, this interpretability evaluation can only be done qualitatively, therefore, lacking objectivity. In future work, we aim to find ways of quantifying the information obtained with the interpretability analysis.

Acknowledgements

This work was financed by the ERDF – European Regional Development Fund through the Operational Programme for Competitiveness and Internationalization - COMPETE 2020 Programme and by National Funds through the Portuguese funding agency, FCT – Fundação para a Ciência e a Tecnologia within project "POCI-01-0145-FEDER-030707", and within the PhD grants "SFRH/BD/137720/2018", "SFRH/BD/139468/2018" and "SFRH/BD/06434/2020".

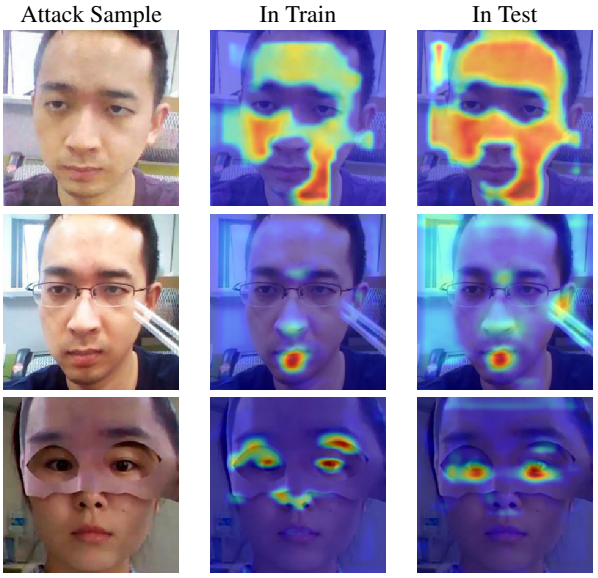


Figure 3: Grad-CAM Explanations for correctly classified attack samples when a subject is in the train set (2nd column) or in the test set (3rd column). Each row corresponds to one specific type of attack, top to bottom: #1, #4, and #7.

References

[1] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140, 2015.

[2] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

[3] Pedro M Ferreira, Ana F Sequeira, Diogo Pernes, Ana Rebelo, and Jaime S Cardoso. Adversarial learning for a robust iris presentation attack detection method against unseen attack presentations. In *2019 International Conference of the Biometrics Special Interest Group (BIOSIG)*, pages 1–7. IEEE, 2019.

[4] Sebastian Lapuschkin, Alexander Binder, Grégoire Montavon, Klaus-Robert Muller, and Wojciech Samek. Analyzing classifiers: Fisher vectors and deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2912–2920, 2016.

[5] Haoliang Li, Wen Li, Hong Cao, Shiqi Wang, Feiyue Huang, and Alex C Kot. Unsupervised domain adaptation for face anti-spoofing. *IEEE Transactions on Information Forensics and Security*, 13(7): 1794–1809, 2018.

[6] Daniel Pérez-Cabo, David Jiménez-Cabello, Artur Costa-Pazo, and Roberto J López-Sastre. Deep anomaly detection for generalized face anti-spoofing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019.

[7] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[8] Ana F Sequeira, Wilson Silva, João Ribeiro Pinto, Tiago Gonçalves, and Jaime S Cardoso. Interpretable biometrics: Should we rethink how presentation attack detection is evaluated? In *2020 8th International Workshop on Biometrics and Forensics (IWBF)*, pages 1–6. IEEE, 2020.

[9] Wilson Silva, Kelwin Fernandes, and Jaime S Cardoso. How to produce complementary explanations using an ensemble model. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2019.

Classification of Not Suitable for Work Images: A Deep Learning Approach for Arquivo.pt

Daniel Bicho[(a)(b)]
daniel.bicho@gmail.com

Artur Ferreira[(b)(c)]
artur.ferreira@isel.pt

Nuno Datia[(b)(d)]
nuno.datia@isel.pt

[(a)] Arquivo.pt, and
[(b)] ISEL, Instituto Superior de Engenharia de Lisboa
Instituto Politécnico de Lisboa
[(c)] Instituto de Telecomunicações

[(d)] NovaLincs, FCT, Universidade Nova de Lisboa

Abstract

Arquivo.pt is a Web Archiving initiative, storing contents preserved from the .pt Web Pages. Among these contents, there are many image files. Some of these images explicitly nudity and pornography, which are offensive for the users, and thus are Not Suitable For Work (NSFW) images. In this paper, we propose a solution to classify NSFW images on Arquivo.pt, using deep learning approaches. We set up a dataset of images with Arquivo.pt data and the ResNet and SqueezeNet models, are evaluated and fine tuned for the NSFW classification task. These models reported an accuracy of 93% and 72%, respectively. After a fine tuning stage, the accuracy of these models improved to 94% and 89%, respectively. This solution is available at `https://arquivo.pt/images.jsp`.

1 Introduction

The collection of portions of the World Wide Web (WWW) to preserve information is named as Web Archiving (WA). This preservation keeps old and historical information available for future use by the general public. Typically, Web Archives resort to Web crawlers, such as Heritrix [10], to collect the web contents. These contents include many resource types, and among them we often find images and videos. There are different WA initiatives, such as the European Commission Historical Archives, `https://ec.europa.eu/historical_archives`, the national top-level domain UK Web Archive, `https://www.webarchive.org.uk/ukwa`, or the Internet Archive, `https://archive.org`. Arquivo.pt, `https://arquivo.pt`, is a WA initiative to preserve the Portuguese *.pt* top-level domain. It provides a research infrastructure, making its contents searchable and publicly available in open access. Arquivo.pt provides a full-text search system and an Image Search Service (ISS) to browse to all its data. This service enables image retrieval capabilities to Arquivo.pt , with an interface in which users can perform queries in natural language and the service retrieves images related to the user query.

One portion of the images stored at Arquivo.pt are Not Suitable For Work (NSFW) for most users, because they contain offensive or explicit images (such as naked persons, violence, and pornography). This may be caused, for instance, by a website that got hacked for Web spam before it was crawled and its contents retrieved. Thus, we need to avoid the exposure to these types of contents, mainly for children and young persons. An example of this NSFW contents retrieved, using the ISS is depicted in Figure 1, in which the query term *angela* was fulfilled on the ISS, and some retrieved results can be considered offensive. In this paper, we propose an approach to filter out nudity/pornography content from the Arquivo.pt resources, through a binary classification task. The remainder of this paper is organized as follows. Section 2 briefly reviews related work. Section 3 presents our approach. The experimental results and some concluding remarks are reported in Section 4.

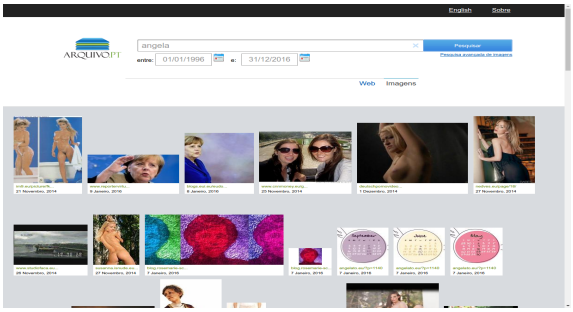


Figure 1: Example of Arquivo.pt problematic content, retrieved with the ISS using the query term *angela*.

2 Not Suitable For Work Image Classification

There are many approaches to the problem of identify NSFW content from images [3, 4, 9, 14]. Among these methods, we can find the first techniques based on skin detection. Another type of techniques are based on Bag-of-Visual-Words (BoVW) and more recently Neural Networks (NN) and Deep Learning (DL) techniques have been proposed.

An automatic system to detect human nudes was present in an image was proposed [4]. It resorts to methods to mark skin-like pixels combined with color and texture properties. These marked regions are then analyzed by a specialized grouper, to group a human figure using geometric constraints on the human structure, followed by classification. The POESIA filter [2], is an open source implementation of a skin-color-based filter. These methods present high false positive rates in images related to beach as well as sports activities.

Another approach is the Bag-of-Visual-Words (BoVW) [3], which extracts, a set of visual features represented as words, setting up a vocabulary vector with the number of occurrences of these visual words representing local image features. Those features usually are derived from detecting keypoints or local descriptors variations. A classifier that uses these representations is then trained to classify the image content.

Recently, Deep Neural Networks (DNN) and more specifically Convolutional Neural Networks (CNN) showed state of the art results, for image recognition tasks. For instance, CNN have been widely used on image recognition tasks [8, 12], for NSFW image classification [13, 15]. Many different CNN architectures have been published with improved accuracy on the standard ImageNet classification challenge [11].

3 Proposed Approach

3.1 Building a dataset

We started by building a dataset of NSFW images and its opposite with 17 655 images, manually labelled from Arquivo.pt with 8 273 labelled as NSFW and 9 382 as SFW, as described in Table 1.

Table 1: Evaluation Dataset.

	SFW	NSFW	Total
Labelled Dataset	9 382	8 273	17 655
Non-Labelled Dataset	-	-	18 626

These images were acquired from Arquivo.pt using two methods. The first method was through the existent Beta Images Indexes, which are Lucene indexes provided by the Solr platform[1]. The second method was through Arquivo.pt Text Search API [1], querying the API to retrieve Web pages, and from those Web pages the images contained were extracted to be manually labelled. On Arquivo.pt the total of images that belong to the NSFW class is much less than the images from the SFW class. The main difficulty at this task is to find enough images from the NSFW class, in order to build a dataset with a significant number of images and with both classes balanced in terms of the number of images. A large number of noisy images were being returned, such as image banners and icons.

[1] `http://lucene.apache.org/solr/`

To reduce the amount of this type of images, only images with resolution above 150x150 pixels were considered in the experimental results reported in Section 4.

3.2 Proposed solution and integration on Arquivo.pt

After building the dataset, we have considered two different topologies of DNN, namely ResNet [7] and SqueezeNet networks [5]. We have addressed the problem as a binary classification task, and thus we used the Cross-Entropy as the loss function [6] and the Stochastic Gradient Descent (SGD) algorithm as the optimizer, using transfer learning. The developed solution to classify image contents as NSFW is integrated in the Arquivo.pt ISS indexing workflow, to extract images and related metadata. The integration is modular, and can be extended by changing the underlying model. It also supports a real time classification, exposed as a Web Service. Figure 2 shows an example of the solution integration, showing a case of a misclassified image by the NSFW classifier (hidden by a gray rectangular box). Figure 3 shows the outcome of the same query, without using the NSFW classifier.

4 Experimental Evaluation and Discussion

The hardware used to evaluate these models is a common laptop with 8 GB RAM, a GeForce GTX 860M as GPU and a Intel(R) Core(TM) i7-4710HQ CPU @ 2.50 GHz. The models were also tested using server class hardware available at Arquivo.pt infrastructure. The server is a Dell PowerEdge R730xd model with 256 GB RAM and an Intel(R) Xeon(R) CPU E5-2620 v3 @ 2.4 GHz. Table 2 reports the best experimental results (accuracy and loss) on the NSFWSqueezeNet model, while Table 3 does the same for the OpenNSFW model.

Table 2: NSFWSqueezeNet fine-tuning accuracy and loss.

Model	4-Fold Accuracy	4-Fold Loss
NSFWSqueezeNet 1 Ep. Aug. 10K	0.88 ± 0.002	0.28 ± 0.004
NSFWSqueezeNet 1 Ep. Aug. 10K	0.89 ± 0.002	0.27 ± 0.006

There is a significant accuracy improvement, from the initial model accuracy of 72% to 89%, after a fine tuning stage in which all the network layers are retrained, using as starting point the network weights from a pre-trained model. The OpenNSFW model is computationally more expensive to train. With the limited hardware available and time constraints, an attempt to improve it was made, freezing all the network layers and retraining only the last fully-connected and the softmax layers. The solver used was also the SGD with the same learning parameters as the model above. The OpenNSFW model provides better results than the SqueezeNet model.

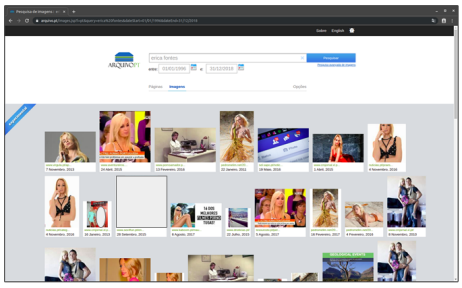


Figure 2: Image filtering interface integration, with query term 'Erica Fontes', with the NSFW classifier. The gray rectangular box highlights NSFW contents which were misclassified.

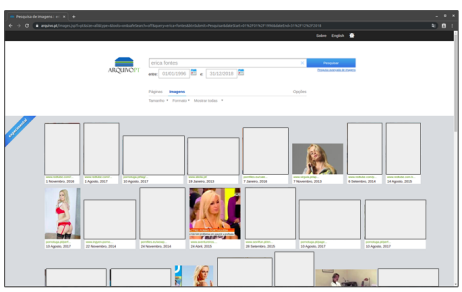


Figure 3: Image filtering interface integration, with query term 'Erica Fontes', without the NSFW classifier. The gray rectangular boxes correspond to NSFW contents (hidden for proper display).

Table 3: OpenNSFW fine-tuning accuracy and loss.

Model	4-Fold Accuracy	4-Fold Loss
OpenNSFW 1 Ep. Aug. 10K	0.92 ± 0.003	0.20 ± 0.006
OpenNSFW 5 Ep. Aug. 10K	0.94 ± 0.004	0.16 ± 0.007

In summary, in this paper we have briefly described a solution that automatically identifies not suitable for work images, which was integrated into Arquivo.pt infrastructure. The solution uses a convolutional neural network to identify this content type and provides the classification result which is used to hide not suitable for work contents, from the retrieved results. The proposed solution is currently available at `https://arquivo.pt/image.jsp`. As future work, the model's accuracy can be improved by building a larger dataset and considering the categorization with more classes.

References

[1] Arquivo.pt. Arquivo.pt API v.0.2 (beta version), https://github.com/arquivo/pwa-technologies/wiki/Arquivo.pt-API-v.0.2-(beta-version), March 2018.

[2] M. Daoudi. POESIA - Filtering Software @ONLINE, January 2018. URL `http://www.poesia-filter.org:80/`.

[3] T. Deselaers, L. Pimenidis, and H. Ney. Bag-of-visual-words models for adult image classification and filtering. In *2008 19th International Conference on Pattern Recognition*, pages 1–4, Dec 2008. doi: 10.1109/ICPR.2008.4761366.

[4] D. Forsyth and M. Fleck. Automatic detection of human nudes. *International Journal of Computer Vision*, 32(1):63–77, Aug 1999. ISSN 1573-1405. doi: 10.1023/A:1008145029462. URL `https://doi.org/10.1023/A:1008145029462`.

[5] F. Iandola, M. Moskewicz, K. Ashraf, S. Hang, W. Dally, and K. Keutzer. SqueezeNet. *arXiv, 1602.07360*, 2016. ISSN 0302-9743. doi: 10.1007/978-3-319-24553-9.

[6] K. Janocha and W. Czarnecki. On loss functions for deep neural networks in classification. *CoRR*, abs/1702.05659, 2017.

[7] H. Kaiming, X. Zhang, S. Ren, and J. Sun. Deep Residual Learning for Image Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. ISBN 978-1-4673-8851-1. doi: 10.1109/CVPR.2016.90.

[8] A. Krizhevsky, I. Sulskever, and G. Hinton. ImageNet Classification with Deep Convolutional Neural Networks. *Advances in Neural Information and Processing Systems (NIPS)*, pages 1–9, 2012.

[9] T. Lindeberg. Scale Invariant Feature Transform. *Scholarpedia*, 7 (5):10491, 2012. doi: 10.4249/scholarpedia.10491.

[10] G. Mohr, M. Stack, I. Ranitovic, D. Avery, and M. Kimpton. Introduction to Heritrix, an archival quality web crawler. In *4th International Web Archiving Workshop (IWAW04)*, Bath, UK, 2004.

[11] Stanford University. ImageNet, http://www.image-net.org/, January 2018.

[12] Y. Sun, B. Xue, M. Zhang, and G. G. Yen. Evolving deep convolutional neural networks for image classification. *IEEE Transactions on Evolutionary Computation*, 24(2):394–407, 2020.

[13] D. Zhelonkin and N. Karpov. Training effective model for real-time detection of nsfw photos and drawings. In *International Conference on Analysis of Images, Social Networks and Texts*, pages 301–312. Springer, 2019.

[14] H. Zheng, M. Daoudi, and B. Jedynak. Blocking Adult Images Based on Statistical Skin Detection. *Electronic Letters on Computer Vision and Image Analysis*, 4(2):1–14, 2004. ISSN 1577-5097. doi: 10.5565/rev/elcvia.78.

[15] R. Zhu, X. Wu, B. Zhu, and L. Song. Application of pornographic images recognition based on depth learning. In *Proceedings of the 2018 International Conference on Information Science and System*, pages 152–155, 2018.

Increasing Road Safety with Machine Learning - A Fatigue and Drowsiness Detection System

António Cerca[(a)]
a40232@alunos.isel.pt
André Lourenço[(a)(b)(c)]
andre.lourenco@isel.pt
Artur Ferreira[(a)(c)]
artur.ferreira@isel.pt

[(a)] ISEL, Instituto Superior de Engenharia de Lisboa
Instituto Politécnico de Lisboa
[(b)] CardioID Technologies
[(c)] Instituto de Telecomunicações

Abstract

In order to make the roads safer both for drivers and pedestrians, there is an increasing interest in monitoring drivers conditions. In this paper, we propose a system that monitors the drivers fatigue and drowsiness, based on both the persons ElectroCardioGram (ECG) signal and the motion of the steering wheel. The acquired data is compressed and transmitted, with a Bluetooth Low Energy profile. A machine learning approach is taken to detect fatigue and drowsiness patterns. The Support Vector Machines classifier proved to achieve the highest accuracy on this task. The low-cost proposed prototype has the ability to warn the driver about his physiological and physical states, thus increasing road safety.

1 Introduction

The driving abilities of a person are affected by two key factors: fatigue and drowsiness. Fatigue is a physical or psychological exhaustion. A person feels fatigued when, for instance, goes to a gymnasium for a reasonable amount of time or when one has solved a large amount of complex problems. Fatigue, usually results from doing the same task repeatedly or in an exhaustive way. Drowsiness is defined as the state before sleep. When someone is drowsy, one requires to sleep, and one's body is fighting to stay awake.

In the past years, we have seen an increasing interest in the development of Advanced Driver Assistance Systems (ADAS), which monitors the vehicle performance and behaviour, as well as the physiological and physical conditions of the driver. These systems resort to accelerometers and other devices to measure the acceleration and other physical quantities. These devices can be placed on the automobile steering wheel to monitor the movements. Moreover, some physiological signals such as electrocardiogram (ECG) [11], can acquired and monitored. The ECG signal can be obtained with the aid of dry-electrodes placed on the vehicle steering wheel. Thus, fatigue and drowsiness detection can be achieved with machine learning algorithms working on these signals. We can identify sleepiness in both the ECG and the steering wheel accelerometer data and to predict if the driver is entering in a state of sleepiness. This detection triggers an alarm to the driver.

The remainder of this paper is organized as follows. Section 2 briefly reviews some concepts on fatigue, drowsiness, and monitoring systems. Section 3 describes our solution. Some experimental results and concluding remarks are reported in Section 4.

2 Monitoring Systems

2.1 Drowsiness scale

The Karolinska Sleepiness Scale (KSS) [9] classifies the drowsiness state with a 10-point Likert scale [4], in which the person classifies his/her sleepiness in periods of 5 minutes. Table 1 describes the KSS scale.

Monitoring systems use sensors and devices to measure parameters for a given purpose. There are two main types of monitoring: direct monitoring and indirect monitoring. Direct monitoring systems deal with physiological signals or with a person behaviour [2]. Indirect monitoring systems interact with the objects controlled by the individual, for example, in an automobile, it is possible to monitor the steering wheel movements, pedal acceleration (gas or break) and sitting position. This kind of monitoring has the advantages of being more robust (usually not influenced by external sources) and more private, since the methods are non-intrusive to the person. Moreover, indirect monitoring systems are easier to use, as compared to direct monitoring systems, on a person that is driving.

2.2 Biometric signals

The electrocardiogram (ECG) signal is the electrical signal that the heart emits [3, 6, 7, 11]. The acquisition of ECG signals can be done in two different ways: using intrusive or non-intrusive methods [1]. Intrusive methods are used in clinical settings where biological signals are extracted using devices placed in the human skin. Non-intrusive methods allow the acquisition of signals with sensors not placed on the person's body, but rather in objects of everyday use. The acquisition of these signals with dry-electrodes is almost involuntarily, without having an impact on the person's daily actions. Figure 1 shows a typical ECG waveform.

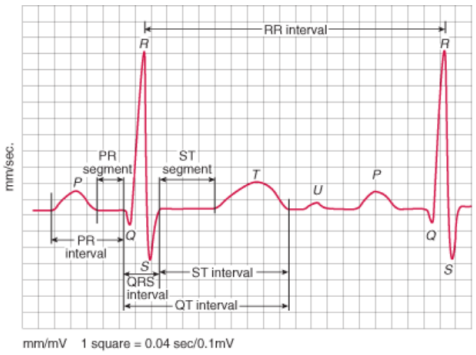


Figure 1: An example of a typical ECG signal [3, 6, 7, 11].

3 Proposed Solution

3.1 Block diagram and prototype placement

The proposed approach is based on the idea that fatigue and drowsiness lead to modifications in the persons biological signals and behaviour. Thus, the monitoring of the fatigue and drowsiness states lead to an adequate approach to warn the driver about his/her state. The acquisition device transmits the data to the gateway and the classification algorithm labels the data and determines if the driver is drowsy or not. When the system determines that the driver is drowsy, the alarm is activated. Figure 2 depicts the block diagram of the proposed system. Our solution is composed by two main parts: the acquisition system, for data collection, preprocessing, and transmission tasks; the gateway solution, for data reception, classification, and alarm activation.

3.2 Acquisition, compression and transmission

Our solution can collect, in a non-intrusive way, the driver ECG signal using dry-electrodes placed in a conductive leather cover (that can fit into

Table 1: The 10-point Karolinska Sleepiness Scale (KSS) [9]

Level	Description
1	Extremely alert
2	Very alert
3	Alert
4	Rather alert
5	Neither alert nor sleepy
6	Some signs of sleepiness
7	Sleepy, but no effort to keep awake
8	Sleepy, but some effort to keep awake
9	Very sleepy, great effort to keep awake, fighting sleep
10	Extremely sleepy, cant keep awake

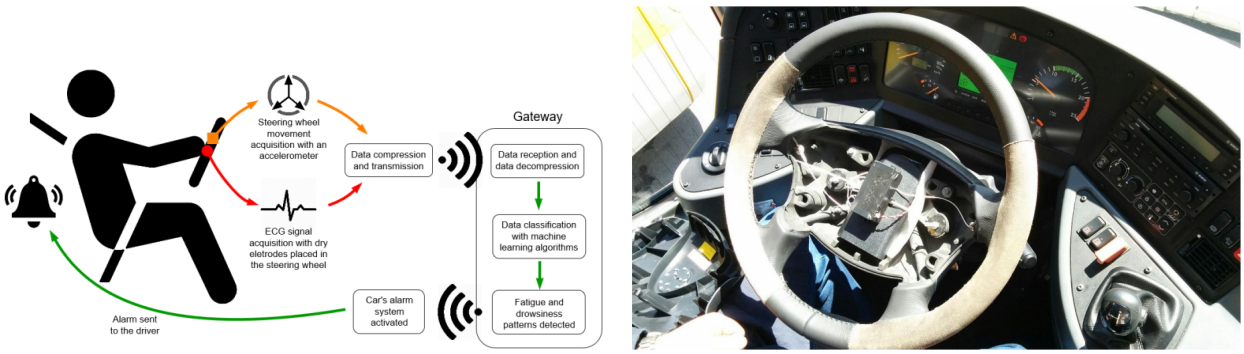


Figure 2: Block diagram of the proposed system (left) and its placement on the automobile (right).

any automobile), and the Steering Wheel Angle (SWA) signal, using an three-axis accelerometer, placed in the centre of the steering-wheel behind the airbag. The dry-electrodes can sense the heartbeat, by its electrical impulses, while the person places the hands on the steering wheel. This electrical continuous signal is converted from analogue to digital with an Analogue-to-Digital Converter (ADC) and the resulting samples are read by a microcontroller. The driver, while moving the steering wheel, causes a variation in each accelerometer axis, and with it, being possible to estimate the rotational angle of the steering wheel. For data compression, we have considered transform-based methods followed by a lossless source coding block [5, 8]. Transform-based methods are the most used techniques to perform lossy encoding of audio and image data. The transform methods are lossless being applied to enable better coefficient quantisation, introducing loss, which results in a lower quality output with high compression ratio. These techniques consist in discarding less significant information, on the quantisation stage, which tends to be irrelevant to the human and (machine) perception of the signal. For transmission, we have considered the Bluetooth Low Energy (BLE) [12] technique, since the devices are battery-powered. BLE allows communications up to 100 meters, in the 2.4 GHz frequency band with rates up to 2 Mbit/s. The current consumption with this technology is around 15 mA.

3.3 Classification

In order to have classifier to work on the data, a feature-based vector must be composed. In the literature, some features were pointed out as being adequate to describe the relationship between ECG or SWA signals with the KSS scales. We have considered sets of 3, 5, and 8 features for the ECG signal, the SWA signal, and both the ECG and SWA signals. For classification purposes we have considered different typical classifiers. Our experimental results have shown that Support Vector Machines (SVM) provide the best results.

4 Experimental Results and Discussion

We have used the dataset provided by the Swedish National Road and Transport Research Institute [1], which contains signals from 18 different people, including ECG and SWA, for the same car and track, in both awake and drowsy states, as well as the KSS values for each data sample. The features from those signals will be the input and the KSS values will be the output to train the classifier. The dataset holds ECG, EEG, and EOG biometric signals, and car movement signals such as velocity, lateral and longitudinal acceleration, Steering Wheel Angle (SWA) and yaw rate. In the experiment, each person was classifying his sleepiness according to the KSS test while driving, adding a KSS value to each data sample. The 9-class output was transformed into a binary classification problem, such that the KSS values above 6 are labelled as a drowsy state [10].

Table 2 reports the experimental results for the classification task, with Linear Regression (LinReg), Logistic Regression (LogReg), Artificial Neural Networks (ANN), using common standard accuracy measures, on the ECG + SWA signals. We have found that it is preferable to use the ECG + SWA signals, as compared to the individual use of the ECG and SWA signals. The SVM classifier, with default parameters, achieves the best results, although it seems to exist some room for improvement. The developed low-cost solution is easy to install on any automobile. It requires no driver cooperation and it achieves interesting results regarding drowsiness detection. As future work, we intend to fine-tune the SVM classifier and to extend this approach to a multi-class classification task.

Table 2: Experimental results for classification of the ECG + SWA signals

Method	Accuracy	Specificity	Recall	Precision	F1-Score
LinReg	0.55	0.58	0.52	0.55	0.50
LogReg	0.55	0.60	0.49	0.55	0.51
ANN	0.54	0.55	0.53 4	0.54	0.51
SVM	**0.62**	**0.56**	**0.68**	**0.61**	**0.64**

[1] https://www.vti.se/en/

Acknowledgements

The authors thank Professor Christer Ahlström, a Senior Researcher at the Swedish National Road and Transport Research for providing us with real data to the classification task.

References

[1] C. Carreiras, A. Lourenço, H. Silva, and A. Fred. Comparative study of medical-grade and off-the-person ECG systems. In *Proc International Workshop on Pervasive Electrocardiography - IWoPE*, 2013.

[2] D. Haupt, P. Honzik, P. Raso, and O. Hyncica. Steering wheel motion analysis for detection of the driver's drowsiness. In *2nd Int. Conf. on Mathematical Models for Engineering Science*, pages 256–261, 2011.

[3] S. Jalaledinne, C. Hutchens, R. Strattan, and W. Coberly. ECG data compression techniques - a unified approach. *IEEE*, 37(4):329 – 343, April 1990.

[4] A. Joshi, S. Kale, S. Chandel, and D. Pal. Likert scale: Explored and explained. *British Journal of Applied Science & Technology*, 7 (4):396–403, 2015.

[5] L. Portolés. Lossles compression of ECG signals - performance analysis in a wireless network, Universitat Politécnica de Catalunya, MSc Thesis, 2009.

[6] Priyanka and I. Saini. Analysis ECG data compression techniques - a survey approach. *International Journal of Emerging Technology and Advanced Engineering*, 3, February 2013.

[7] C. Saritha, V. Sukanya, and Y. Murthy. ECG signal analysis using wavelet transforms. *Bulg. J. Phys.*, 35:68–77, 2008.

[8] K. Sayood. *Introduction to data compression*. Morgan Kaufmann, 4th edition, March 2012.

[9] A. Shahid, K. Wilkinson, S. Marcu, and C. Shapiro. *Karolinska Sleepiness Scale (KSS) - STOP, THAT and One Hundred Other Sleep Scales*. Springer, New York, NY, 2011.

[10] C. Silveira. Drivers fatigue state monitoring using physiological signals. Master's thesis, Universidade do Porto, 2017.

[11] U. Tiwary and B. Gohel. Automated risk identification of myocardial infarction using relative frequency band and coefficient (RFBC) features from ECG. *The Open Biomedical Engineering Journal*, 4: 217–222, 2010.

[12] K. Townsend. Introduction to bluetooth low energy, Adafruit Industries, 2018.

Radiomic analysis of brain MRI: A case study in Autism Spectrum Disorder

Joana Soeiro[1]
joana.soeiro@ua.pt
Lília Dias[1]
liliadias@ua.pt
Augusto Silva[2]
augusto.silva@ua.pt
Ana Tomé[2]
ana@ua.pt

[1] DFIS, Univ. Aveiro
Aveiro, Portugal

[2] IEETA, DETI, Univ. Aveiro
Aveiro, Portugal

Abstract

This work is intended to study the relationship between the morphology and texture of different brain structures and the presence of the disease called Autism. For that, it is proposed a radiomic analysis of the brain structures, amygdala and hippocampus, that will be performed by feature extraction and further analysis of the changes that occur in patients diagnosed with autism. Through the analysis of the textural features, it is expected the discovery of potential biomarkers that when combined with the morphological information, will possibly assist the diagnosis and a better understanding of Autism.

1 Introduction

Autism Spectrum Disorder (ASD) is a disease that develops in children, usually between 2/3 years old, being characterized by motor, cognitive and emotional difficulties. The causes of this disease are not yet fully understood, which has led to an increase in the study of this disorder. Recent studies indicated that the development of autism can cause structural and functional changes in the amygdala and the hippocampus, which are brain structures that are responsible for controlling some emotional and cognitive behaviours. For this reason, an effort has been made to understand the changes in those structures with the development of autism [1, 2]. One of the imaging approaches to visualize the changes that occur in brain structures is the Magnetic Resonance Imaging (MRI). Through this technique, 3D brain images can be acquired enabling subcortical quantitative analysis with focus on the amygdala and hippocampus. Once the images are obtained, features can be extracted from them, allowing a posterior analysis [1]. The extraction and mining of a high number of quantitative features in an automated and high-throughput way is called radiomics [1, 3]. Regional variations of texture in an image can be analysed in order to provide information about the brain structures in study. Its analysis is done through techniques dedicated to the quantification of the spatial variation of the gray tones of the image [4]. To the best of our knowledge, some of the existent studies in this field are contradictive, for example some conclude that in patients diagnosed with autism, there is an increase in the volume of the mentioned structures, while others indicate the opposite [5, 6].

The present work has the goal to find biomarkers capable of diagnose ASD. In order to achieve it, two approaches were followed: a first and general approach, by exploring the mean values of the features extracted from the MRI images, and a second approach done through a radiomics analysis of classification based on a similar study performed by Chaddad *et al* [1].

2 Methods

The data used was obtained from ABIDE I database, that provide MRI images gathered from several medical institutions around the world. A sample of 48 patients was used for this study, in which 24 were apparently healthy, being used as control and 24 were diagnosed with autism. In what concerns the age, only 10 had less than 15 years in which 5 had autism. Regarding the gender, 37 patients were male and 11 female. It is important to notice that the number of patients with different genders and ages was already determined in the samples taken from ABIDE I, not being determined within this project. The procedure was divided in three steps: (1) brain images segmentation, (2) feature extraction and (3) feature analysis.

2.1 Brain Images Segmentation

The MRI images obtained from the database were submitted to the VolBrain online segmentation workflow, capable of segmenting the brain into several structures (e.g. amygdala, hippocampus, cerebellum, cortex, etc) [7].

2.2 Feature extraction

The segmented images obtained from VolBrain were submitted to LifeX, a software capable of extracting morphologic (volume) and texture features taken from the gray-level co-occurrence matrix (GLCM) and histograms of each structure of the brain [8]. Six features obtained from the GLCM were extracted: energy, entropy, contrast, correlation, homogeneity and dissimilarity. The energy measures the homogeneity by the sum of squares of entries in the GLCM , the entropy represents de disorganization of the gray levels, the contrast symbolize local variations of the image intensity, the correlation represents the linear gray-level reliance between pixels and the homogeneity detects the similarity between the gray levels [9]. From the histograms, four features were extracted: skewness, kurtosis, entropy and energy. The skewness is a feature that gives information about the symmetry of the histogram and the kurtosis gives information about its flatness [9].

2.3 Feature analysis

Two types of feature analysis were performed. The first analysis carried out was a general exploratory approach, which may not be indicative of anything, since it was done through the observation of the mean values of the features. The main goal of this approach was to have a first look on the behaviour of the features, being analysed the quantitative difference between the average value of the same feature between cases and controls and also between age groups (<15 and > 15 years).

The second approach was done through case classification based on the feature space. The case classification using features both from the amygdala and hippocampus, were performed with three different models: support vector machine (SVM), neural networks (NN) and random forest (RF). In the non-linear SVM the radial basis function was used as kernel and in the linear SVM the linear function was used as kernel. Furthermore, in the SVM and NN models, the data had to be separated in training and testing and two different methods were used: the k-fold and the leave one out (LOO). Since that the RF classifier already does the separation of the data into training and testing, it was not necessary to apply the LOO or the k-fold methods, as in the other classifiers used. The k-fold method consists of dividing the total data set into $\boldsymbol{k}$ subsets of the same size and, from there, a subset is used for testing and the remaining $\boldsymbol{k}$-1 are used to estimate the parameters. This process is performed $\boldsymbol{k}$ times by alternating the test subset. The LOO method is similar to the previous one but instead of dividing the data into folds, it leaves only one example for test and the rest is for training. Thus, the process is carried out $\boldsymbol{N}$ times, equal to the number of sample sets [10]. The performance of the models was evaluated by their accuracy, obtained from the confusing matrix. The classification was performed in three different types of data: (a) features extracted only from the hippocampus, (b) features extracted only from the amygdala and (c) features extracted from both amygdala and hippocampus. Another relevant characteristic of the RF classifier is the possibility of measuring the importance of each feature on the classification process, allowing to rank them. This ranking was performed in 15 runs, allowing the obtaining of the TOP features in the classification of each data (a, b and c).

3 Results and discussion

Figure 1 presents the images obtained from the LifeX software, where it is possible to see the structures in study that resulted from the segmentation.

Figure 1: Frontal, sagittal and inferior images of the brain, highlighting the structures in study.

From the first exploratory analysis of the features, the average values of the volume of the two structures of the brain, separated by age and diagnostic were obtained and are represented in Table 1. It is possible to verify that in patients under the age of 15 years, both structures have a lower volume when compared to individuals of the same age range without autism. The same happens when comparing the volume of the hippocampus between individuals with autism and controls, over 15 years. However, in this age group it appears that the volume of the amygdala in individuals with autism is higher than that of controls.

Table 1: Volume (cm^3) of both studied structures of the patients (with and without autism) that participate in this study.

Age	Patients with autism		Controls	
	Amygdala	**Hippocampus**	**Amygdala**	**Hippocampus**
<15	2.09	9.94	2.47	10.21
>15	1.56	9.50	1.02	9.72

The same exploratory analysis was carried out with the textural features. Regarding the textural characteristics obtained from the histograms, it was possible to verify that the hippocampus in patients with autism presented higher values of skewness, kurtosis and entropy and lower energy than in patients without this disorder. The same features taken from the amygdala of patients with autism presented lower values of skewness and kurtosis and higher values of entropy and energy. The same analysis made of the textural features taken from the GLCM in the hippocampus showed, in patients with autism, lower values of homogeneity, energy, contrast, correlation and dissimilarity and higher values of entropy. In the amygdala, the same analysis showed higher values of homogeneity, correlation, entropy and dissimilarity and lower energy and contrast in patients with autism. The same analysis was done separating the sample in age groups (<15 and > 15 years) however, it was not found a general rule to the behaviour of the features. It should be noticed that this analysis is only exploratory, and the quantity of patients used should be larger in order to obtain more conclusive results. Also, the number of patients under 15 years used for this study was significantly less than the patients above 15 years, giving only a general idea of the differences in the age groups.

Through the second approach of the analysis of the features, the accuracy of each classifier used were obtained, being possible to evaluate each one of them. The results obtained are shown in Table 2.

Table 2: Accuracy values of each model used to classify the data used, for both LOO and k-fold split methods.

	Hippocampus		Amygdala		Hippocampus + amygdala	
	LOO	**kf**	**LOO**	**kf**	**LOO**	**kf**
SVM linear	0.500	0.575	0.625	0.575	0.625	0.642
SVM non linear	0.500	0.592	0.563	0.625	0.500	0.500
NN	0.604	0.592	0.500	0.567	0.688	0.633
RF	0.958		0.958		0.980	

It should be noted that a greater accuracy was obtained with the RF model in all the data used, both with the LOO method and with the k-fold to separate the data in training and test datasets. The LOO method is computationally expensive, which in this case does not represent a problem because of the small size of the sample [10]. However, if the study was projected in a bigger sample, preference would be given to the k-fold method. The dataset containing characteristics of the amygdala and hippocampus had better results compared to the data containing only characteristics of the amygdala or the hippocampus. This may indicate that by analysing the structures together, it is possible to obtain better results in classifying patients with autism using this models.

For each set of classified data, the TOP features were obtained after 15 runs, as mentioned before, and are represented in Table 3. Firstly, it should be noted that the features extracted from the GLCM are those that allow a better classification of the structures, since they appear with more frequency compared to the ones extracted from the histograms. It is possible to verify that the volume of the amygdala, entropy, homogeneity and dissimilarity are the most common characteristics of the 3 data sets. This way, it is possible to conclude that these features can be used as biomarkers to identify autism.

Table 3: TOP features obtained through the RF classification model for each data set used.

Amygdala	**Hippocampus**	**Hippocampus + amygdala**
Age	Homogeneity (GLCM)	Amygdala volume
Volume	Energy (GLCM)	Amygdala entropy (GLCM)
Contrast (GLCM)	Entropy (GLCM)	Hippocampus homogeneity (GLCM)
Correlation (GLCM)	Dissimilarity (GLCM)	Hippocampus entropy (GLCM)
Entropy (GLCM)		
Dissimilarity (GLCM)		

4 Conclusion

This work intended to analyse MRI brain images, in particular the hippocampus and amygdala structures, in order to contribute to the diagnose and understanding of the ASD. This was possible through a radiomic analysis of the MRI images, extracting texture and morphologic features and analysing them by two different approaches. This way, it was possible to compare those feature between patients with and without autism. The first approach allowed the general understanding of the behaviour of the features and the second approach allowed us to achieve the features that presented more potential to become biomarkers in the diagnosis of ASD. In order to achieve this in the second approach, several feature classification models were used and evaluated through their accuracy, allowing to achieve the conclusion that the RF model presented better results in classifying the patients in cases or controls. It was also with this model, that was possible to obtain the features with more potential to be used as biomarkers in the diagnosis of autism.

References

[1] Chaddad et al, "Hippocampus and amygdala radiomic biomarkers for the study of autism Ima spectrum disorder" BMC Neurosci (2017) 18:52 DOI 10.1186/s12868-017-0373-0

[2] Xu, Qinfang, et al. "Abnormal development pattern of the amygdala and hippocampus from youth to adolescent with autism." Journal of Clinical Neuroscience (2020).

[3] Lambin, Philippe, et al. "Radiomics: extracting more information from medical images using advanced feature analysis." European journal of cancer 48.4 (2012): 441-446.

[4] Gibbs, Peter, and Lindsay W. Turnbull. "Textural analysis of contrast-enhanced MR images of the breast." Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine 50.1 (2003): 92-98.

[5] Schumann, Cynthia Mills, et al. "The amygdala is enlarged in children but not adolescents with autism; the hippocampus is enlarged at all ages." Journal of neuroscience 24.28 (2004): 6392-6401.

[6] Aylward, Elizabeth H., et al. "MRI volumes of amygdala and hippocampus in non–mentally retarded autistic adolescents and adults." Neurology 53.9 (1999): 2145-2145.

[7] José V. Manjón and Pierrick Coupe. volBrain: an online MRI brain volumetry system. Frontiers in Neuroinformatics. 2016, https://volbrain.upv.es/

[8] C Nioche, F Orlhac, S Boughdad, S Reuzé, J Goya-Outi, C Robert, C Pellot-Barakat, M Soussan, F Frouin, and I Buvat. LIFEx: a freeware for radiomic feature calculation in multimodality imaging to accelerate advances in the characterization of tumor heterogeneity. Cancer Research 2018; 78(16):4786-4789, www.lifexsoft.org

[9] Chaki, Jyotismita, and Nilanjan Dey. Texture Feature Extraction Techniques for Image Recognition. Springer Singapore, 2020.

[10] S. Raschka, "Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning", University of Wisconsin–Madison Department of Statistics, November 2018.

Sentinel-2 Image Scene Classification over Alentejo Region Farmland

Kashyap Raiyani[1]
kshyp@uevora.pt
Teresa Gonçalves[1]
tcg@uevora.pt
Luis Rato[1,2]
lmr@uevora.pt
Pedro Salgueiro[1]
pds@uevora.pt
José R. Marques da Silva[3,4]
jmsilva@uevora.pt

[1] Departmento de Informática, Universidade de Évora, Portugal
[2] CIMA, Universidade de Évora, Portugal
[3] MED, Universidade de Évora, Portugal
[4] Agroinsider Lda., Évora, Portugal

Abstract

Given the wide-ranging farmland area, optical satellite images of farms are used to develop maps that reflect land dynamics and its behavior over different time frames, crops, and regions on various environmental conditions. In this regard, it is essential to identify and remove atmospheric distorted images to further prevent misleading information, since their presence severely restrict the use of optical satellite images for forecasting harvest dates, yield estimation, and manufacturing control in agriculture systems. These atmospheric distortions are frequent due to cloud, shadow, snow, and water cover over farmland. In this work, we developed a method to identify distortion covering images of corn crop farmland situated in the Alentejo Region of Portugal. The results are compared with the state-of-the-art (SOTA) Sen2Cor algorithm of the European Space Agency. Further, experimental results show that the developed image scene classifier model outperforms Sen2Cor by 10% in F1-measure.

1 Introduction

Agriculture in Europe has witnessed a substantial change after the creation of the Common Agriculture Policy (CAP)[1] in 1962. As a result, Food security [6] is ensured in most parts of Europe but the estimated global population growth 7 billion to 9 billion by 2050 [2] posses the challenge of producing more food [12]. The way to address this challenge is to rely on science and technology for possible answers.

Over the last few decades, many new technologies have been developed for or adapted to, agricultural use. Examples of these include low-cost positioning systems such as the Global Navigation Satellite System (GNSS) or the Geographic Information Systems (GIS), sensors mounted on agricultural machinery, geophysical sensors aimed at measuring soil properties, low-cost remote sensing techniques, and reliable devices to store, process and exchange/share information [3, 10]. Together, these new technologies have produced a large amount of affordable, high resolution information and have led to the development of site-specific agricultural management that is often termed Precision Agriculture.

There are many aspects related to Precision Agriculture and this work aims at investigating Sentinel-2 satellite images (or known as product) to gain information across different parcel/region and time. Resulting, a time data-series that takes land (usage) properties as input and outputs land dynamic which will provide information about environmental (such as soil, water and, weather) impact on the land and crop growth.

The existence of optical distortion such as clouds, shadows, snow, and water over land can mask true surface reflection resulting in false land information restricting the use of satellite images. To identify this distortion, sate-of-the-art (SOTA) Sen2Cor image scene classifier could be used. Sen2Cor is an algorithm whose main purpose is to correct single-date Sentinel-2 Level-1C products from the effects of the atmosphere and deliver a Level-2A surface reflectance product [7]. Level-2A (L2A) output consists of a Scene Classification (SCL) image with seven classes: Cirrus, Shadow, Snow, Water, Vegetation, Soil, and Cloud with low, mid, and high probability.

This document reports the work developed within the scope of the NIIAA (Núcleo de Investigação em Inteligência Artificial em Agricultura), project co-promoted by the company Agroinsider[1]. In this regard, we created a Sentinel-2 image scene classifier, and used the developed classifier over the corn parcel images to recognize atmospheric distortion.

[1] https://ec.europa.eu/info/food-farming-fisheries/key-policies/common-agricultural-policy/cap-glance_en

2 Developed Work

The health of plants can be determined by their biophysical parameters and can be measured by spectral information gathered using remote sensing. The physiological changes (due to crop stress) lead to a change in the spectral reflection/emission characteristics [8]. This observation of the stress factor during crop growth using, for example the Normalized Difference Vegetation Index (NDVI) [11] is a necessary stage to know the probable loss of production. NDVI values are affected by multiple factors such as available soil moisture, date of planting, air temperature, day length, and soil condition [9].

2.1 Study Area

With the help of Agroinsider, we acquired 170 (5 days apart) Sentinel-2 images from 05-01-2017 to 03-08-2019 of ten corn parcels from Alentejo region between (37°56'29.13" N , 8°22'21.95" W) and (37°55'32.44" N, 8°21'02.23" W) coordinates. Figure 1 shows the corresponding 2D image of the ten corn parcels (referred as parcel-1 to parcel-10 onwards).

Figure 1: Ten Corn Parcels from Alentejo Region.

Figure 2 shows the mean NDVI Value from 05-01-2017 to 03-08-2019 for parcel-1[2]. In it, the presence of atmospheric disturbance can be observed as sudden dips in the NDVI values, supported by the fact that it is not possible to lose crop growth and regain it within a range of 5 days (the observation cycle time). To calculated mean NDVI, for each point in the parcel, NDVI was calculated using equation 1, and the overall sum value was divided by the total number of points. Here, NIR means Near Infra-Red (Band 8) and RED is Band 4.

$$NDVI = (NIR - RED)/(NIR + RED) \quad (1)$$

2.2 Scene Classification and Results

Holstein [4] created a database of manually labeled Sentinel-2 spectra. The database consists of images acquired over the entire globe and comprises 6.6 million points from 60 different products classified into six classes as clear-sky, cloud, cirrus, shadow, snow, and water. The database is described by 4 attributes: *product_id*, *latitude*, *longitude* and *class*. To build a classifier, we extended this database adding corresponding Sentinel-2 13 bands values and, for comparison purposes, Sen2Cor scene classification. The final structure of the database is detailed in Table 1.

Instead of using standard train_test_split from Scikit-Learn library [5], we selected 59 products for training, and 1 for testing. The main reason to split the dataset in this way was to make sure that the knowledge about

[2] The same can be replicated to rest of parcels.

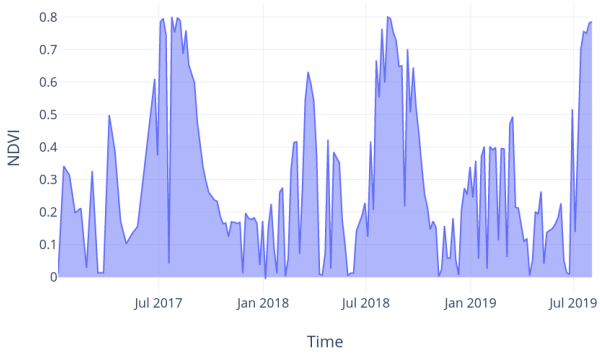


Figure 2: Mean NDVI Value for parcel-1 from 05-01-2017 to 03-08-2019.

Header	Column Value
Product ID	1 Column (78 character string)
Coordinates	4 Columns (latitude, longitude, east and, north)
Bands	13 Columns (Band 1 to 12 and 8A)
Tagged Class	1 Column (Manual tagged class value)
Sen2cor - SCL	1 Column (Scene classification class value)

Table 1: Structure of Final Dataset.

a region is not essential to classify that region. This reasoning enables us to pose the following question: will the system be able to classify it with good performance a new, non seen product? To evaluate this, it would be interesting to pick a complete region as a test set while all the rest of the points compose the training set. We replicated this procedure for each of the 60 products (use 1 for test and the rest 59 for train). We present the average F1 results. Equation 2 calculates the average $F1$ value (over 60 products) for each class where $F1_p$ is the $F1$ value of the particular class within the product p. N_p is the number of points of the class within the product p, T is the total number of points of the class for all products and $p \in (1,60)$ is the number of products.

$$F1 = \sum_{p=1}^{60} \left(F1_p \times N_p\right) \div T \;\; with \;\; T = \sum_{p=1}^{60} N_p \tag{2}$$

We used the Scikit-Learn library implementation of Decision Tree (DT), Random Forest (RF) and Extreme Trees (ET) algorithms. The obtained results were compared with the Sen2Cor algorithm. Table 2 details the results. These results show an F1 average value of 76.77% over all classes (using Extreme Trees), an improvement over 10% when compared to Sen2Cor F1 average value of 66.40%.

Class	DT	RF	ET	Sen2Cor	Support
Clear-sky	63.29	72.3	**74.16**	64.96	1694454 (25.56%)
Water	63.81	73.4	76.69	**80.73**	1071426 (16.16%)
Shadow	53.98	**63.96**	61.45	50.57	991393 (14.96%)
Cirrus	47.58	**56.63**	42.97	24.08	956623 (14.43%)
Cloud	65.25	75.08	**75.33**	75.04	1031819 (15.57%)
Snow	74.67	84.90	**87.00**	61.40	882763 (13.32%)
$F1_{avg}$	67.95	76.43	**76.77**	66.40	6628478 (100%)

Table 2: F1 values of ML algorithms and Sen2Cor.

Using the developed Extreme Tree model, the new, unseen optical images (with 13 bands) of the ten parcels mentioned in Subsection 2.1 were classified as no atmospheric disturbance image (clear-sky) or image with disturbance (cloud, shadow, snow, and water coverage). Here, each point within the parcel was classified using the model ET model built, resulting in a value between 0 if all points were classified as clear sky and 1 when all points were classified as atmospheric disturbance. Figure 3 presents the calculated disturbance over dates 14-06-2017 to 01-12-2017, with red line for the ET model and blue line mean NDVI. These results sync with sudden dips of the NDVI values supporting the claim of the presence of atmospheric disturbance in the optical image.

After analyzing Figure 3 closely, the authors would like to state that 'NDVI value is not the sole parameter to find disturbance'. This claim is supported by Figure 3 as on 08, 13, and 18 Aug'17, the mean NDVI ranges from 0.78 to 0.68 (a drop) to 0.76 but the value of atmospheric disturbance remains 0.0.

3 Conclusion

From our experiment results (Table 2), RF and ET are comparatively providing equivalent results and outperforming Sen2Cor by 10% F1 measure for image scene classification over a specific dataset composed by 6.6M

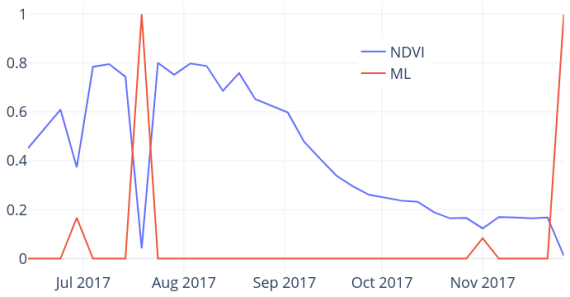


Figure 3: Parcel-1: Mean NDVI and Atmospheric Disturbance Identification by ML (over dates 14-06-2017 to 01-12-2017).

entries acquired from 60 different products. Further, the results in Figure 3 support our claim: the ML model presented in this work is applicable as a base tool to identify the existence of clouds, shadows, snow, and water coverage over agriculture farmland images acquired using Sentinel 2 optical satellite. As a result, classified parcel images will help to prevent false surface reflectance information and allow the use of selected optical images for forecasting harvest dates, yield estimation, and manufacturing control. Given that the train ML model is over 60 different products acquired over the entire globe comprises 6.6 million points, the author expects similar results of identifying atmospheric disturbances over different crops.

As future work, we would like to: (1) manually label individual data points for each parcel as (0 or 1) atmospheric disturbance and, (2) compare the performance of the ML method to Sen2Cor over each parcel.

Funding

This work was supported by the NIIAA (Núcleo de Investigação em Inteligência Artificial em Agricultura) project, Alentejo 2020 program (reference ALT20-03-0247-FEDER-036981).

References

[1] Agroinsider an agricultural consulting company. `https://www.agroinsider.com/`. Accessed: 04 09 2020.

[2] Nikos Alexandratos and Jelle Bruinsma. World agriculture towards 2030/2050: the 2012 revision. 2012.

[3] Gibbons. Turning a farm art into science-an overview of precision farming. *URL: http://www. precisionfarming. com*, 2000.

[4] André Hollstein, Karl Segl, Luis Guanter, Maximilian Brell, and Marta Enesco. Ready-to-use methods for the detection of clouds, cirrus, snow, shadow, water and clear sky pixels in sentinel-2 msi images. *Remote Sensing*, 8(8):666, 2016.

[5] Oliver Kramer. Scikit-learn. In *Machine learning for evolution strategies*, pages 45–53. Springer, 2016.

[6] Deepak Kumar and Prasanta Kalita. Reducing postharvest losses during storage of grain crops to strengthen food security in developing countries. *Foods*, 6(1):8, 2017.

[7] Jérôme Louis, Vincent Debaecker, Bringfried Pflug, Magdalena Main-Knorn, Jakub Bieniarz, Uwe Mueller-Wilm, Enrico Cadau, and Ferran Gascon. Sentinel-2 sen2cor: L2a processor for users. In *Proceedings Living Planet Symposium 2016*, pages 1–8. Spacebooks Online, 2016.

[8] ARR Menon. Remote sensing application in agriculture and forestry. *T Sabu, T., Vinod. T R., M Subramonia, I., C. Bhaskaran., and B. Ambat*, pages 222–235, 2012.

[9] M Duane Nellis, Kevin P Price, and Donald Rundquist. Remote sensing of cropland agriculture. *The SAGE handbook of remote sensing*, 1:368–380, 2009.

[10] Francis J Pierce and Peter Nowak. Aspects of precision agriculture. In *Advances in agronomy*, volume 67, pages 1–85. Elsevier, 1999.

[11] JW Rouse, RH Haas, JA Schell, and DW Deering. Monitoring vegetation systems in the great plains with erts. *NASA special publication*, 351:309, 1974.

[12] Pablo J Zarco-Tejada, N Hubbard, and P Loudjani. Precision agriculture: An opportunity for eu farmers—potential support with the cap 2014-2020. *Joint Research Centre (JRC) of the European Commission*, 2014.

Deep learning to automate the assessment of cultural ecosystem services from social media data

Ana Sofia Cardoso[1]
up201804335@fc.up.pt
Ana Sofia Vaz[2]
sofia.linovaz@gmail.com
Francesco Renna[1]
frarenna@dcc.fc.up.pt

[1]Instituto de Telecomunicações, Faculdade de Ciências da Universidade do Porto
[2]Andalusian Inter-University Institute for Earth System Research (IISTA-CEAMA), University of Granada & Research Centre in Biodiversity and Genetic Resources (CIBIO-InBIO), University of Porto

Abstract

Cultural ecosystem services (CES) result from the interactions between humans and nature, contributing to people's physical and mental well-being. Most social media content analyses considered in the context of CES are based on the manual classification of photos or texts shared by social media users. Inevitably, the manual classification of big photographic data is too time consuming and costly, particularly when it comes to large study areas and audiences. In this work we studied automated image classification techniques using deep learning approaches to address CES.

1 Introduction

Nowadays, computer science and related fields have been highly invested in the use and combination of methods that incorporate social media analytics [1]. Social media platforms represent a very significant fraction of all the available digital data, constituting an efficient method to collect big data that provide information on people's interactions with each other and with their environment [2]. Fast improvements in computational power and data storage capacity during the last years have motivated the emergent fields of Digital Conservation, iEcology and conservation culturomics [3]. These disciplinary fields refer to the use of digital (big) data and technology to understand human-nature interactions and to provide evidence in favour of nature conservation and of the sustainable management of ecosystems [4]. Among these human-nature interactions are cultural ecosystem services (CES), which constitute the non-material benefits that people can experience from nature, such as recreation and ecotourism, as well as those pertaining to spiritual, religious, aesthetic or heritage values, among others [5].

An approach that combines different data from social media with advanced analytics, besides spatial analysis, remains underexplored in the context of CES assessment. Thus, the investment in methods that can identify features of ecosystems and nature through the content analysis of shared photos (or text), can constitute an asset to support the evaluation of CES, particularly, related to aesthetics and recreation or ecotourism [6]. Lee *et al.*, for example, proposed a method for analysing large amounts of social media photographs, as well as to derive indicators of socio-cultural usage of landscapes, through cluster detection with Convolutional Neural Networks (CNNs) [7]. This project aims to develop an automated classification of social media photographs that can be useful for CES evaluation and for providing innovative solutions to the scientific community. Specifically, this study aims to answer the following questions: (1) can deep learning algorithms be developed to support an automated classification of social media photographs in the context of CES? and (2) how can those algorithms and models be improved so as to promote statistically reliable image classifications? To achieve this, deep learning algorithms are developed and tested, more specifically CNNs and transfer learning strategies are applied to the classification of digital photographs of the "Peneda-Gerês" protected area (Northern Portugal) obtained from the social media platforms Flickr and Wikiloc.

2 Methods

2.1 Image classification methodology

We performed a classification of the content of photographs from the protected area "Peneda-Gerês" (Northern Portugal), that were withdrawn from the Flickr and Wikiloc social media platforms, specifying a time window of 2003-2017 (1778 images in total). This classification was based on "Nature" and "Human" labels (Figure 1). To achieve that, two different CNNs architectures were implemented, the VGG16 and the ResNet152, in order to verify the most appropriate and suitable for our study.

The proposed image classification methods were evaluated over the dataset using a 5-fold-cross validation method, following the literature and taking into account the computational resources and the running time. The considered performance metrics (accuracy, sensitivity, specificity, and F1-score) were computed as the mean of the performance metrics obtained over the 5 different folds. During training, in each of the 5 folds, 10% of the training data was retained to perform model validation, in order to determine the training parameters that guaranteed the highest accuracy over the validation set.

Since we are coping with a small dataset, in order to improve the generalization of the model and avoid the overfitting, transfer learning and data augmentation schemes were considered.

Figure 1: Examples of images belonging to the Nature and Human labels. a) Nature, b) Human.

2.2 CNN architectures and transfer learning

The VGG16 and ResNet152 were the chosen CNNs architectures. For both CNN architectures, three different sets of weights were considered: (1) weights obtained by training over the dataset "Places365", (2) weights obtained by training over the database "ImageNet" and (3) weights obtained by training the networks from scratch.

The Places365 dataset is the latest subset of the database Places, comprising around 1.8 million scene photographs of different places, labelled with 365 scene semantic categories, including photographs with similar elements to the ones under study. The ImageNet database constitutes a large-scale hierarchical image database, that has several applications in the broadest areas, comprising more than 14 million cleanly annotated images spread over around 21,000 categories. Both databases were selected due to their freely available online resources (weights and models).

Regarding the details of the transfer learning strategy implemented, all the convolutional layers were kept frozen when training over our dataset, while the remaining 3 (for VGG16) and 1 (for ResNet152) fully connected layers were trained with our dataset. Moreover, for both architectures, an additional dense layer with 128 units and a rectifier linear unit activation function was also included (to allow better fit of the model/network to the classification task) before the output layer, which was modified in order to have 2 units.

Regarding the training details, both networks were trained using the Adam optimizer. For VGG16, the best performance was verified when considering a learning rate of 0.000001 while, for ResNet152, it was 0.0001 the most accurate learning rate. Also, it was observed that, for VGG16, the model accuracy and loss had fully converged after 50 epochs, having been decided, because of that, to use only 50 epochs to build the VGG16 model, as well as the ResNet152 model, due to computing resource management.

2.3 Data augmentation

Regarding data augmentation, 5 transformations (including horizontal flip, width shift, height shift and zoom) were implemented individually for each of the images in the training set. The images in the validation set were not included in this process, in order to avoid biased results. The total number of transformations applied to each photograph (5 per image) was selected taking into account the overall running time of the algorithm, as well as the available computational memory.

3 Results

3.1Nature vs. Human classification

When comparing the two transfer learning scenarios and the weights obtained by training only over our dataset (Figure 1), it was observed that, ImageNet had, overall, a higher accuracy for the two architectures under study (86.11 vs 87.18), followed by Places365 and weights trained only with our dataset, with the exception of Places365 in VGG16, that resulted in an equally high accuracy (87.01). Also, it was verified that, for Places365, VGG16 had a better performance when compared to ResNet152 (87.01 vs 86.00), while for the remaining scenarios, ResNet152 model was more accurate than the one for VGG16.

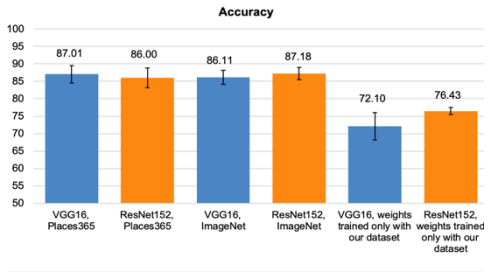


Figure 1: Accuracy of the VGG16 and ResNet152 model performance for the two transfer learning scenarios and the weights from scratch.

Considering sensitivity (Figure 2), it was verified that ImageNet had, overall, better results for the two architectures under study (86.71 and 86.78), followed by Places365 and weights trained only with our dataset, with the exception of Places365 in VGG16, that resulted in a higher sensitivity value (88.48). Likewise, it was observed that ResNet152 had slightly finer sensitivity results when compared to VGG16, except for Places365, where VGG16 showed the best result (88.48 vs 83.40).

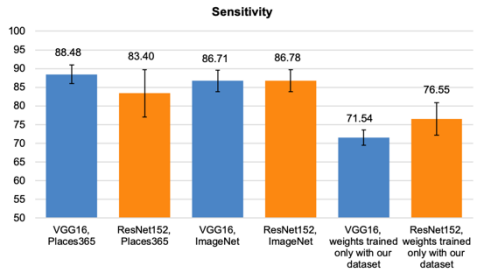


Figure 2: Sensitivity of the VGG16 and ResNet152 model performance for the two transfer learning scenarios and the weights from scratch.

For specificity (Figure 3), it was observed that Places365 had finer specificity results for the two architectures under study (85.54 and 88.46), followed by ImageNet and weights trained only with our dataset. Similarly, it was verified that ResNet152 had better specificity results when compared to VGG16, for all the scenarios under study.

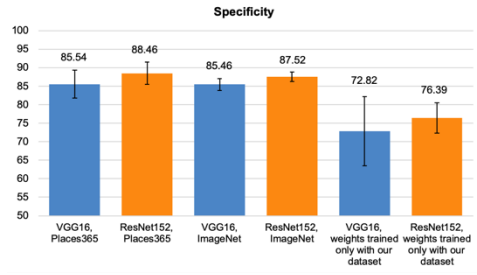


Figure 3: Specificity of the VGG16 and ResNet152 model performance for the two transfer learning scenarios and the weights from scratch.

Considering the F1-score (Figure 4), it was verified that ImageNet had slightly better F1-score results for the two architectures under study (86.53 and 87.44), followed by Places365 and weights trained only with our dataset. Also, it was observed that ResNet152 had finer F1-score results when compared to VGG16, except for Places365, where VGG16 showed the best result (87.53 vs 85.89).

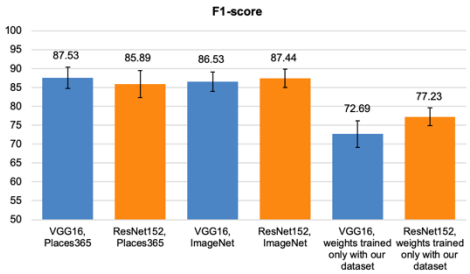


Figure 4: F1-score of the VGG16 and ResNet152 model performance for the two transfer learning scenarios and the weights from scratch.

4 Discussion and Conclusions

When comparing the two considered transfer learning scenarios and the weights obtained by training only over our dataset, it was expected that the model implemented with the Places365 weights would have a finer performance than the other two (with ImageNet weights and weights trained only with our dataset), since all the photographs contained in this dataset are exclusively related with landscapes and places in general, constituting the database that most resembles our dataset. Perhaps surprisingly, this was not the case for both VGG16 and ResNet152, as ImageNet was undoubtedly the database where the two transfer learning scenarios achieved better results. A possible explanation for this behavior can reside in the observation that deep learning models achieve more accurate results when trained in the presence of large datasets. In fact, ImageNet, by containing a larger number of photographs (more than 14 million) than Places365 (around 1.8 million), has led to a better performance of the model. Also, ImageNet contains a greater diversity of images that seems to contribute to a better generalization of the model.

The results showed that deep learning methods can offer significant contributions to assist in CES evaluation. Future work will focus on the improvement of the robustness of these models against scarcely labeled data via the use of semi-supervised approaches by leveraging autoencoder architectures and generative adversarial networks.

Acknowledgements

This work is funded by FCT/MCTES through national funds and when applicable co-funded EU funds under the project UIDB/50008/2020.

References

[1] Sherren Kate et al. Conservation culturomics should include images and a wider range of scholars. *Frontiers in Ecology and the Environment*, 2017, 15.6: 289-290. doi: 10.1002/fee.1507.

[2] Di Minin et al. Prospects and challenges for social media data in conservation science. *Frontiers in Environmental Science*, 2015, 3: 63. doi: 10.3389/fenvs.2015.00063.

[3] Jarić Ivan et al. iEcology: Harnessing Large Online Resources to Generate Ecological Insights. *Trends in Ecology & Evolution*, 2020. doi: 10.1016/j.tree.2020.03.003.

[4] Toivonen Tuuli et al. Social media data for conservation science: a methodological overview. *Biological Conservation*, 2019, 233: 298-315. doi: 10.1016/j.biocon.2019.01.023.

[5] Assessment Millennium Ecosystem et al. *Ecosystems and human well-being* (Vol. 5). United States of America: Island press, 2005. Doi:

[6] Richards Daniel R.; Tunçer, Bige. Using image recognition to automate assessment of cultural ecosystem services from social media photographs. *Ecosystem services*, 2018, 31: 318- 325. doi: 10.1016/j.ecoser.2017.09.004.

[7] Lee Heera et al. Mapping cultural ecosystem services 2.0–Potential and shortcomings from unlabeled crowd sourced images. *Ecological Indicators*, 2019, 96: 505-515. doi: 10.1016/j.ecolind.2018.08.035.

IHC Classification in Breast Cancer H&E Slides with a Weakly-Supervised Approach

Sara P. Oliveira[1,2] (sara.i.oliveira@inesctec.pt)
João Ribeiro Pinto[1,2] (joao.t.pinto@inesctec.pt)
Tiago Gonçalves[1,2] (tiago.f.goncalves@inesctec.pt)
Hélder P. Oliveira[1,3] (helder.f.oliveira@inesctec.pt)
Jaime S. Cardoso[2,1] (jaime.cardoso@inesctec.pt)

[1] INESC TEC, Porto, Portugal
[2] FEUP, Porto, Portugal
[3] FCUP, Porto, Portugal

Abstract

Human epidermal growth factor receptor 2 (HER2) evaluation commonly requires immunohistochemistry tests on breast cancer tissue, in addition to the standard haematoxylin and eosin (H&E) staining. Additional costs and time spent on further testing might be avoided if HER2 overexpression could be inferred from H&E slides, as a preliminary indication of the IHC result. We propose a framework that separately processes H&E slide tiles and outputs an IHC label for the whole slide. The network was trained on slides from the HER2 Scoring Contest dataset (HER2SC) and tested on two disjoint subsets of slides from the HER2SC database and the TCGA-TCIA-BRCA (BRCA) collection. The proposed method attained 83.3% classification accuracy on the HER2SC test set and 53.8% on the BRCA test set. Although further efforts should be devoted to achieving improved performance, the obtained results suggest that it is possible to perform HER2 overexpression classification on H&E tissue slides.

1 Introduction

Breast cancer (BCa) is the most commonly diagnosed cancer and the leading cause of cancer-related deaths among women worldwide. However, over the most recent years, despite the increasing incidence trends, the mortality rate has significantly decreased. Among other factors, this results from better treatment strategies that can be delineated from the assessment of histopathological characteristics [1, 2].

The analysis of tissue sections of cancer specimens obtained by biopsy commonly starts with haematoxylin and eosin (H&E) staining, which is usually followed by immunohistochemistry (IHC), a more advanced staining technique used to highlight specific protein receptors, such as the HER2 [3]. In fact, the overexpression of HER2 is observed in 10%–20%[4] of BCa cases and has been associated with aggressive clinical behaviour and poor prognosis [5]. However, these cases have a better response to targeted therapies and consequent improvements in healing and overall survival [5].

The current guidelines [6], revised by the American Society of Clinical Oncology/College of American Pathologists (ASCO/CAP), in 2018, indicate the following scoring criteria for HER2 IHC:

- IHC 0+: no staining or incomplete, faint/barely perceptible membrane staining in 10% of tumour cells or less;
- IHC 1+: incomplete, faint/barely perceptible membrane staining in more than 10% of tumour cells;
- IHC 2+: weak to moderate complete membrane staining in more than 10% of tumour cells;
- IHC 3+: circumferential, complete, intense membrane staining in more than 10% of tumour cells.

Moreover, cases scoring 0+ or 1+ are classified as HER2 negative, while cases with a score of 3+ are classified as HER2 positive. Cases with score 2+ are classified as equivocal and are further assessed by *in situ* hybridization (ISH), to test for gene amplification [6].

Despite the efficiency of IHC and ISH, the additional cost and time spent on these tests might be avoided if all the information needed to infer the HER2 status could be extracted only from H&E slide, as a preliminary indication of the IHC result. However, to the extent of our knowledge, the task of predicting HER2 status on H&E slides has not yet been addressed in the literature, except for a recent challenge[1].

[1] ECDP2020 HEROHE Challenge: `https://ecdp2020.grand-challenge.org`

2 Methodology

The proposed method (Fig. 1) comprises a CNN, pre-trained for the task of HER2 scoring of IHC tiles. The pre-trained parameters are then transferred to the task of HER2 status prediction on H&E tiles, to provide the network with some knowledge of the tissue structures' appearance. Individual tile scores are then combined in a single label for the whole slide.

2.1 Data Preprocessing

For the IHC slides of classes 2+ and 3+, the preprocessing begins with automatic tissue segmentation with Otsu's thresholding obtaining the regions with more intense staining, that correspond to the HER overexpression areas. For slides of classes 0+ and 1+, the segmentation consists of simple removal of pixels with the greatest HSV value intensity, corresponding to background pixels, which do not contain essential information to the problem. These processes, which are performed at 32× downsampled slides, return the masks used in tile extraction. Tiles with size 256×256 are extracted from the slide with original dimensions (without downsampling), provided they are completely within the mask region. These tiles are converted from RGB to HSL colour space, of which only the lightness channel is used. Each tile inherits the class from the respective slide (examples in Fig. 2a–d), turning the learning task into a weakly-supervised problem.

According to the ASCO/CAP guidelines for IHC evaluation, the diagnosis is performed based only on the tumoral region of the slides. Hence, the preprocessing of H&E slides begins with an automatic invasive tissue segmentation with the HASHI method [10, 11]. The segmentation mask is then used to generate H&E tiles (example in Fig. 2e), extracted and processed according to the abovementioned methodology.

2.2 IHC Tile Scoring & H&E Slide Classification

The CNN architecture for the IHC tile scoring consists of 4 convolutional layers (16, 32, 64 and 128 filters, respectively, with ReLU activation). The first layer has a 5×5 kernel, while the remaining have 3×3 kernels. Each convolutional layer is followed by a pooling layer (a max-pooling function without overlap, with kernel 2×2). The network is topped with three fully-connected layers, with 1024, 256, and 4 units, respectively. The first two have ReLU activation, while the third is followed by softmax activation for the output of probabilities for each class.

The network parameters pre-trained with IHC tiles were used as initialization for HER2 status classification on H&E tiles. To achieve a single prediction per tile instead of four, as it was initially trained for on the IHC setting, a soft-argmax activation [12] replaces the softmax activation.

The output scores are then sorted from 3+ to 0+ and the tiles corresponding to the 15% highest ones are selected for the aggregation process. This percentage was chosen to limit the information given to the aggregation network, while still including and barely exceeding the reference 10% of tumour area considered in the HER2 scoring guidelines.

The score aggregation is performed by a multilayer perceptron (MLP), composed of 4 layers, with 256, 128, 64, and 2 neurons, respectively. All layers are followed by ReLU activation and the last one is followed by softmax. Since the input dimension M of the MLP is fixed (we set $M = 300$ to limit memory cost), for images where 15% of the number of tiles exceeds M, they are downsampled to M using evenly distributed tile selection. In cases where 15% of the number of tiles is lower than M, tiles are extracted with overlap, to guarantee that M tiles can be selected. The MLP will process these M HER2 scores and output a single HER2 status label for the respective slide.

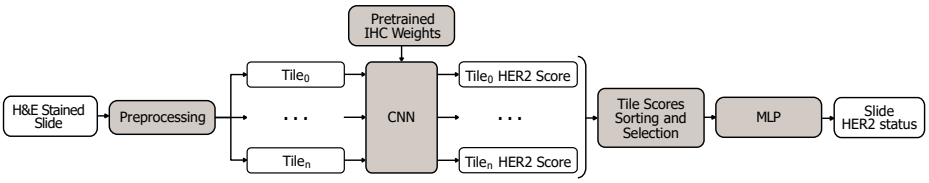
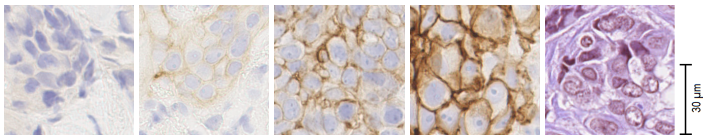


Figure 1: The proposed approach for weakly-supervised HER2 status classification on BCa H&E slides.



(**a**) (**b**) (**c**) (**d**) (**e**)

Figure 2: Tile examples extracted from IHC 0+ (**a**), IHC 1+ (**b**), IHC 2+ (**c**), IHC 3+ (**d**), H&E (**e**) slides. Examples extracted from [7, 8, 9].

3 Data & Training Details

The dataset is composed of subsets of slides from two public datasets: the HER2 Scoring Contest (HER2SC) training set [7] and the TCGA-TCIA-BRCA (BRCA) collection [8, 9]. The HER2SC training set (with available labelling) comprises slides of 52 cases of invasive BCa stained with both IHC and H&E. The subset from the BRCA dataset includes 54 H&E slides. All slides have the same original resolution and are weakly annotated with HER2 status (negative/positive) and score (0+, 1+, 2+, 3+), obtained from the corresponding histopathological reports. The training and validation sets, used for the IHC model parameter tuning and optimization, have 40 and 12 slides, respectively, corresponding to 7591 tiles per class for training (30,364 tiles total) and 624 tiles per class for validation (2496 tiles total), to keep a class balance.

The hyperparameters used during training were empirically set to maximize performance. The CNN model for IHC tile scoring was randomly initialized and trained using the Adaptive Moment Estimation (Adam) optimizer (learning rate of 1×10^{-5}), to minimize a cross-entropy loss function, during 200 epochs, with mini-batches of 128 tiles. The soft-argmax used a parameter $\beta = 1000$. The aggregation MLP was also trained using the Adam optimizer, with learning rate of 10^{-5} for 150 epochs and mini-batches of 1 slide (consisting of soft-argmax scores of the respective 300 tiles), saving the best considering validation accuracy.

4 Results and Discussion

After training, the IHC model offered 76.8% accuracy. This indicates that the model was able to adequately discriminate against the IHC tiles between the four classes.

On the HER2SC test set, this method achieved a weighted accuracy of 83.3% and a F1-score of 86.7% (see Table 1). Despite the small size of this test set, the proposed method was able to correctly classify all positive slides and only misclassify one negative sample as positive. In this context, one might consider this a desirable behaviour, as false positives are less impactful than false negatives.

Table 1: H&E HER2 status classification results of the proposed method.

	Accuracy	**F1-score**	**Precision**	**Recall**
HER2SC	83.3%	86.7%	89.6%	87.5%
BRCA	53.8%	21.5%	81.2%	31.5%

When tested on the BRCA test set, this method achieved a weighted accuracy of 53.8% and a F1-score of 21.5% (see Table 1). The method retains the behaviour presented in HER2SC, preferring to err on the side of false positives than the alternative. On the other hand, the performance metrics on BRCA differ considerably from those obtained on HER2SC. While the method was trained on HER2SC data, which is expected similar to the test data, the slides of the BRCA have a greater extent of tissue, generating more tiles per image and impacting the distribution of the scores, which may influence the method's behaviour.

The other shortcomings of the method appear to be related to the invasive tumour segmentation and the tile scoring network, which could be improved with additional data and more accurate ground truth. With these additional efforts, the proposed method could offer robust weakly-supervised HER2 classification without IHC information.

5 Conclusions

In this work, a framework is proposed for the weakly supervised classification of HER2 overexpression status on H&E BCa slides. The proposed approach integrates a CNN trained for HER2 scoring of individual H&E tiles, initialized with the network parameters pre-trained with data from IHC images. The objective of this initialization is to transfer some domain knowledge to the final training. The individual scores are aggregated on a single prediction per slide, returning the HER2 status label.

The evaluation results in single-database (HER2SC) and cross-database (BRCA) settings show the potential of the proposed method in standard and more challenging situations, indicating that it is possible to accurately infer BCa HER2 status solely from H&E slides.

Despite these results, further efforts should be devoted to performance improvement. Firstly, the training of the tile HER2 scoring CNN and the aggregation MLP could be integrated into a single optimization process. On the other hand, the aggregation of individual scores could use tile locations to take spatial consistency into account. Finally, the knowledge embedded in the networks through the pre-trained parameters could be better seized if input H&E tiles could be previously converted into IHC, for example, using generative adversarial models.

Acknowledgements This work was partially funded by the Project "TAMI: Transparent Artificial Medical Intelligence" (NORTE-01-0247-FEDER-045905), co-financed by ERDF, European Regional Fund through the Operational Program for Competitiveness and Internationalisation (COMPETE 2020), the North Portugal Regional Operational Program (NORTE 2020) and by the Portuguese Foundation for Science and Technology (FCT), under the CMU-Portugal International Partnership, and also the FCT PhD grants "SFRH/BD/139108/2018", "SFRH/BD/137720/2018" and "SFRH/BD/06434/2020".

References

[1] American Cancer Society. Breast Cancer Facts & Figures 2017–2018. Available online: `http://bit.ly/acs_bcff_1718`.

[2] Gandomkar, Z.; Brennan, P.; Mello-Thoms, C. Computer-based image analysis in breast pathology. *J. Pathol. Inform.* **2016**, *7*.

[3] Veta, M.; Pluim, J.P.W.; van Diest, P.J.; Viergever, M.A. Breast Cancer Histopathology Image Analysis: A Review. *IEEE Trans. Biomed. Eng.* **2014**.

[4] American Society of Clinical Oncology (ASCO). Breast Cancer Guide. 2005–2020. Available online: `http://bit.ly/asco_bcg`.

[5] Rakha, E.A. *et al*. Updated UK Recommendations for HER2 assessment in breast cancer. *J. Clin. Pathol.* **2015**, *68*, 93–99.

[6] Wolff, A.C. *et al*. Human Epidermal Growth Factor Receptor 2 Testing in Breast Cancer: American Society of Clinical Oncology/College of American Pathologists Clinical Practice Guideline Focused Update. *J. Clin. Oncol.* **2018**, *36*, 2105–2122.

[7] Qaiser, T. *et al*. HER2 challenge contest: A detailed assessment of automated HER2 scoring algorithms in whole slide images of breast cancer tissues. *Histopathology* **2018**, *72*, 227–238.

[8] Clark, K. *et al*. The Cancer Imaging Archive (TCIA): Maintaining and Operating a Public Information Repository. *J. Digit. Imaging* **2013**, *26*, 1045-1057.

[9] Lingle, W. *et al*. Radiology Data from The Cancer Genome Atlas Breast Invasive Carcinoma [TCGA-BRCA] collection. *Cancer Imaging Arch.* **2016**.

[10] Cruz-Roa, A. *et al*. High-throughput adaptive sampling for whole-slide histopathology image analysis (HASHI) via convolutional neural networks: Application to invasive breast cancer detection. *PLoS ONE* **2018**, *13*, 1–23.

[11] Cruz-Roa, A. *et al*. Accurate and reproducible invasive breast cancer detection in whole-slide images: A Deep Learning approach for quantifying tumor extent. *Sci. Rep.* **2017**, *7*.

[12] Honari, S. *et al*. Improving landmark localization with semi-supervised learning. In CVPR, 19–21 June 2018; pp. 1546–1555.

Mood Estimation Based on Facial Expressions and Postures

Daniel Canedo
danielduartecanedo@ua.pt
António J. R. Neves
an@ua.pt

IEETA/DETI
Universidade de Aveiro
Aveiro, 3810-193, Portugal

Abstract

This paper presents a mood estimation algorithm based on facial expressions and postures using Computer Vision and Deep Learning. This algorithm consists in two well-known modalities within Computer Vision: facial expression recognition and pose estimation. Such algorithm can be useful in a wide range of applications that may benefit from feedback regarding the mood of a user. A specific application that estimates the mood of a speaker during a speech was used for testing the developed software. The obtained results are preliminary, although promising in terms of accuracy.

1 Introduction

Facial expressions, postures and gestures are visible indicators that depict someone's feelings. However, estimating these indicators using Computer Vision still raises many challenges. For instance, most facial expression recognition datasets were built around posed facial expressions and controlled scenarios. As studied in [1], it is difficult to translate the accurate results in controlled environments into real world scenarios. A common strategy to face this problem is to perform a meticulous data preprocessing. Normalizing the data usually leads to a significant improvement on the accuracy of Machine Learning models. Postures and gestures are important means to express emotions and to communicate behavioral intentions. Although some studies seem to indicate that postures and gestures contribute equally for emotion recognition [2], they are not being explored as much as facial expressions within this research problem.

The main contribution of this work was to build a multimodal algorithm capable of estimating mood. An example of an application for such algorithm is also presented: estimating the mood of a speaker. This could also be used, for example, to diagnose mental disorders, to monitor risky driving behaviors, to improve marketing strategies based on the estimated people's reaction and to improve human-computer interaction.

2 Related Work

Deep Learning based algorithms have been really popular in the last few years. This convergence towards Deep Learning is correlated with overall better results in several areas, and Computer Vision is no exception. Regarding facial expression recognition, several recent papers claim to have achieved around 98% accuracy in controlled environment datasets using Deep Learning solutions. However, this high accuracy is still not translatable to real world scenarios. Since most facial expression recognition datasets are built around controlled environments and the subjects are asked to pose certain facial expressions, the samples are somewhat artificial. This discrepancy can be understood in a recent paper [3]: the proposed solution attained 98.90% accuracy when testing on the Extended Cohn-Kanade (CK+) dataset [4], but it only obtained 55.27% accuracy on the Static Facial Expression in the Wild (SFEW) dataset [5]. The CK+ dataset was built around a controlled environment and posed facial expressions, while the SFEW dataset was built around uncontrolled environments. Head pose variation, different lighting conditions and posed facial expressions are the main contributors to such discrepancy. However, there is an Emotion Recognition in the Wild Challenge (EmotiW) that has been stimulating solutions for uncontrolled environments in facial expression recognition. The recent winners of this challenge are mainly building multimodal classifiers and performing face and intensity normalization. They have pushed the state-of-the-art accuracy on facial expression recognition in uncontrolled environments to 63.39% [6].

Regarding pose estimation, there are several Deep Learning solutions that are able to accurately return keypoints corresponding to the associated body parts. With these keypoints and their association through time, it is possible to extract relevant information regarding posture and gestures. The state-of-the-art pose estimation algorithms' accuracy ranges from 69% to 80%, reflecting some unsolved challenges of pose estimation: occlusions and body parts association. PoseNet [7], which was the used model for this work, was trained on a ResNet and a MobileNet. The ResNet model has a higher accuracy, but its deep architecture is not ideal for real time applications. On the other hand, the MobileNet model is smaller, providing faster predictions but with less accuracy. In the interest of reducing the processing time, the MobileNet version was considered for this work. When PoseNet processes an image, what is returned is a heatmap along with offset vectors that can be decoded to find high confidence areas in the image, resulting in 17 keypoints.

3 Proposed Approach

Since real time performance was one of the goals of this work, a simple CNN was designed for facial expression recognition. Figure 1 illustrates the proposed CNN architecture.

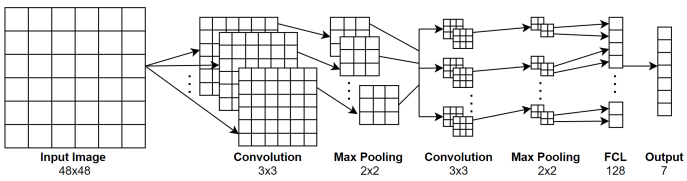


Figure 1: Proposed CNN architecture for facial expression recognition.

This CNN architecture receives as input a 48 × 48 grayscale image since facial expression recognition models tend to perform better for this resolution and higher [8]. The rest of the architecture is standard, consisting of two convolutional layers, two max-pooling layers, three batch normalization layers and one fully connected layer with a dropout layer. The CK+ dataset was used for training. Before the training step, the dataset was pre-processed by applying rotation correction, cropping, intensity normalization, histogram equalization and smoothing, respectively. Finally, the CNN was trained with the Adam optimizer. Class weights were calculated to deal with the unbalanced data for each class. The batch size was set to 32 and the training data was shuffled in each epoch. The training step was done for 100 epochs and the weights that presented the best validation accuracy were saved. It is possible to observe in Figure 2 that the proposed CNN achieved 93% validation accuracy and did not overfit.

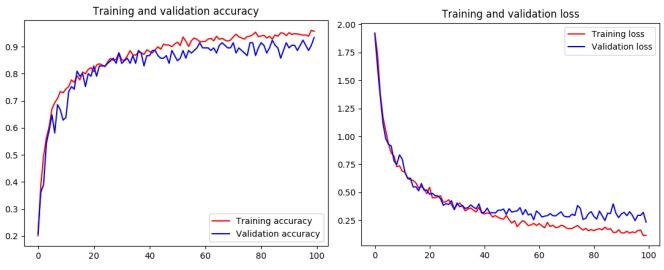


Figure 2: Training step results of the facial expression recognition model.

Regarding pose estimation, it was implemented the pre-trained MobileNet model of PoseNet as mentioned in Section 2. As a potential application for the developed software, estimating the mood of a speaker was considered. Being confident during a speech is often correlated with a good understanding of the topic. A study suggests that having an expansive body posture is often correlated with dominance, power and confidence, while not having an expansive body posture often reflects low self-esteem and apprehension [9]. Therefore, in this work, the expansiveness of a speaker is estimated from the keypoints returned from PoseNet.

It is possible to estimate the expansiveness of a speaker by calculating a ratio between the occupied area [10] and the minimum area that the speaker could be occupying. It can be calculated as follows:

$$A_{min} = |K_{Ymax} - E_{Ymin}| \times |S_{Xmax} - S_{Xmin}| \quad (1)$$

$$A_{current} = |K_{Ymax} - K_{Ymin}| \times |K_{Xmax} - K_{Xmin}| \quad (2)$$

$$A_{ratio} = \frac{A_{current}}{A_{min}} \quad (3)$$

Where E represents the eyes keypoints, S represents the shoulders keypoints, K represents the minimum and maximum keypoints and A represents the area of the bounding box. The minimum area ratio is 1 and the maximum area ratio was truncated to 5. Regarding the facial expression recognition model, it returns one of the six basic emotions (anger, disgust, fear, happiness, sadness, surprise) or the neutral expression.

4 Results and Discussion

A 1-minute segment of a speech given by Professor António J. R. Neves in TEDxAveiro 2019 was used for testing the developed software. When processing the segment with the facial expression recognition model, it was observed that the facial motion of the speaker when he was giving the speech, mainly mouth movement and head pose variation, contributed to some false positives. During the whole segment, the speaker presented a neutral expression, however the facial expression recognition model only detected that expression 50% of the times. This confirms the challenge of uncontrolled environments in facial expression recognition discussed in Section 2. Figure 3 illustrates some false positives triggered by the mouth of the speaker combined with different head poses.

Figure 3: False positives of the facial expression recognition model. From left to right: anger, disgust, fear, happiness, sadness and surprise.

Regarding pose estimation, the segment was successfully processed with PoseNet, which returned the necessary keypoints for estimating the expansiveness of the speaker. Figure 4 illustrates an example of a processed frame.

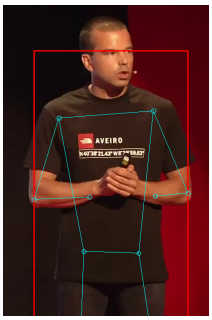


Figure 4: Processed frame from the TEDxAveiro speech segment.

The body lines represent the body parts detected by PoseNet and their keypoints, while the bounding box represents the current area occupied by the speaker, as discussed in Section 3. It can be observed that the bounding box is properly drawn, taking into consideration the horizontal extremities and the vertical extremities: K_{Ymin} is the Y-coordinate of the eyes, K_{Ymax} is the Y-coordinate of the legs, K_{Xmin} is the X-coordinate of the left elbow and K_{Xmax} is the X-coordinate of the right elbow (see Equation 2 from Section 3). During the whole segment, the speaker's posture was the same as Figure 4, which was not expansive. The calculated expansiveness using Equation 3 from Section 3 was 1.59.

Since the minimum expansiveness is 1 and the maximum expansiveness value is 5, it is possible to adapt the facial expression recognition output to the pose estimation output. Table 1 attempts to adapt the facial expression categories to numeric values and Table 2 shows the fusion between the facial expression and expansiveness values with proposed labels.

Category	Value
Negative (anger, disgust, fear, sadness)	1
Neutral (neutral, surprise)	3
Positive (happiness)	5

Table 1: Numeric values of the facial expression categories.

Fusion	Label
1	Anxious
3	Comfortable
5	Confident

Table 2: Fusion between the two modalities and their labels.

Since the calculated expansiveness was 1.59 and the estimated facial expressions were 56.5% neutral, 43% negative and 0.5% positive, the mood of the speaker can be calculated through the following Equation:

$$\textbf{Mood} = \frac{Expansiveness + (Negative + Neutral \times 3 + Positive \times 5)}{2} \quad (4)$$

Using Equation 4, **the estimated mood was 1.87, which is somewhere between anxious and comfortable** (see Table 2). This value is reasonable since the speaker revealed that he was nervous and anxious about the speech, but at the same time he was comfortable since he is an expert on the topic.

The two explored modalities for mood estimation are promising, however in order to increase the trustworthiness of the developed software, it is necessary to improve the facial expression recognition model in uncontrolled environments, as well as adding more relevant modalities, such as tone of voice and movement.

References

[1] Daniel Canedo and António JR Neves. Facial expression recognition using computer vision: A systematic review. *Applied Sciences*, 9(21):4678, 2019.

[2] Beatrice de Gelder, AW De Borst, and R Watson. The perception of emotion in body expressions. *Wiley Interdisciplinary Reviews: Cognitive Science*, 6(2):149–158, 2015.

[3] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. From facial expression recognition to interpersonal relation prediction. *International Journal of Computer Vision*, 126(5):550–569, 2018.

[4] Patrick Lucey, Jeffrey F Cohn, Takeo Kanade, Jason Saragih, Zara Ambadar, and Iain Matthews. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *2010 ieee computer society conference on computer vision and pattern recognition-workshops*, pages 94–101. IEEE, 2010.

[5] Abhinav Dhall, OV Ramana Murthy, Roland Goecke, Jyoti Joshi, and Tom Gedeon. Video and image based emotion recognition challenges in the wild: Emotiw 2015. In *Proceedings of the 2015 ACM on international conference on multimodal interaction*, pages 423–426, 2015.

[6] Abhinav Dhall. Emotiw 2019: Automatic emotion, engagement and cohesion prediction tasks. In *2019 International Conference on Multimodal Interaction*, pages 546–550, 2019.

[7] D Oved, I Alvarado, and A Gallo. Real-time human pose estimation in the browser with tensorflow. js. *TensorFlow Medium, May*, 2018.

[8] Chun Fui Liew and Takehisa Yairi. Facial expression recognition and analysis: a comparison study of feature descriptors. *IPSJ transactions on computer vision and applications*, 7:104–120, 2015.

[9] Dana R Carney, Amy JC Cuddy, and Andy J Yap. Power posing: Brief nonverbal displays affect neuroendocrine levels and risk tolerance. *Psychological science*, 21(10):1363–1368, 2010.

[10] Fábio Barros, Ângelo Conde, Sandra C Soares, António JR Neves, and Samuel Silva. Understanding public speakers' performance: First contributions to support a computational approach. In *International Conference on Image Analysis and Recognition*, pages 343–355. Springer, 2020.

Segmentation of fetus brain MRI based on K-nn algorithm

F. J. Fernandes Oliveira
fjfoliveira.99@gmail.com
Paulo Salgado
psal@utad.pt
T-P Azevedo Perdicoúlis
tazevedo@utad.pt

Dep. Engenharias, ECT
Universidade de Trás-os-Montes e Alto Douro
5000-715 Vila Real, Portugal

Abstract

The segmentation of MRI of the fetal brain has been emerging as a clinical tool to detect abnormalities during the development of the fetus. Since the brain is still in development, a mixture of regions of white matter, grey matter and transition structures that are related to brain growth are still associated with it. In this work, two versions of the K-nearest neighbour algorithm are proposed as the core method for the recognition of different regions of images; the first one is a refinement of the standard algorithm and the second a reinforcing iterative version of the same method. Both versions are used to identify 3 *apriori* selected regions — the brain, the intracranial and the remaining part of the fetus body. The effectiveness of the method has been demonstrated in a MR image segmentation that was first pre-processed with digital filters for feature extraction. Contour filters have also been applied to the same image. The results obtained with the proposed segmentation procedure showed better performance than other traditional methods.

1 Introduction

In the image processing realm, image segmentation (IS) refers to the process of dividing a digital image into multiple regions (sets of pixels), to simplify its representation and facilitate analysis. As a result, a set of regions, or contours, is extracted from the image, where every pixel in the same region has similar characteristics, such as color, intensity, texture, or continuity. In biomedical engineering, the combination of magnetic resonance imaging (MRI) with computational techniques warrants the development of high interest tools for analysis of brain imaging to assist neuro-scientists and general practitioners in the early diagnosis. MRI of the human fetus is emerging as a clinical tool for early detection of brain abnormalities due to its ability to evaluate morphometric measurements of a brain in development and promises a range of new quantitative biomarkers to be used in the clinical evaluation of pregnancy. It is factual that fetus MRI may be corrupted by noise or blurred by the movement of the fetus during the examination. As reported in [4], the anatomy of the developing fetal brain is significantly idiosyncratic, in terms of both geometric and underlying tissue morphology, as it consists of a mixture of white matter, gray matter, and transitory structures related to brain growth. The segmentation of the fetus has been extensively studied by several authors over the years and several methods are described for this type of study. In [5], the construction of an atlas with 10 isotropic images is proposed as a first method. A second method is the construction of a simple atlas based on a segmentation that aims to find the transformation between the atlas and the target low-resolution images. Following a similar methodology, [1] put labels on the atlas to specify different structures of the brain, since these are usually used as a paradigm for automatic segmentation algorithms. In this paper, the fetal brain is segmented through neighbouring characteristics using the K-nn algorithm [2]. It distinguishes different regions automatically as well as structures that are present in the acquired MR images. In our case study, these regions are the brain, the uterus, and partial body. The remainder of the work is organised as follows: Image acquisition and pre-processing are explained in Section 2. The K-nn Algorithm is described in Section 3. The application of the K-nn algorithm and respective results can be found in Section 4. Section 5 concludes the paper and withdraws some directions for future work.

2 Image acquisition and pre-processing

Digital image processing (DIP) refers to the manipulation of digital images through processing methods able to make it more clear or removing noise and other artifacts. Sometimes, DIP is also used to highlight certain characteristics as contours or textures. In this work, test images are downloaded from [3], where one can find several kinds of fetus MRI as the test image shown in Fig. 1. Once the image is uploaded, three square regions of the image are identified as sample sub-images — the brain, the intracranial space and the partial fetal body — whose respective pixels are taken as training data, belonging to the training set (TS). In this work, DIP algorithms are used to provide some local features of the image to the modified K-nn algorithm, which it trying to recognise regions of the image with similar characteristics. The algorithm input are the data features of a circular region around every pixel and the output are the results of two categorical filters: (1) low-pass filters: the mean, the median and the variance; (2) edge or contours operators: Sobel, Canny, Prewitt and the "Laplacian of the Gaussian Operator". In the processing of the image of Fig. 1, low-pass filters were used to measure the mean properties of a local region of the image, since they produce a smoothed image as result. In particular, median filters can reduce image noise without blurring the image contours. This type of filter is specially suited for removing impulsive noise that appears in limited regions of the image. Additionally, the variance filter provides for the information variation contained in the local image region, which is important for distinguishing textures and some kind of patterns. The variance is calculated around each pixel as:

Figure 1: Selected atlas image. The yellow boxes outline the local regions: brain, intracranial space and partial fetal body.

$$Var_{ij} = \frac{1}{n} \sum_{(r,s)\in\mathbb{R}_{ij}} \left(I_{rs} - \bar{I}_{ij}\right)^2, \tag{1}$$

where I_{rs} is the intensity of pixel–(r,s) of the image contained in vicinity $\mathbb{R}_{ij}$ and $\bar{I}_{ij}$ is the mean value of the intensity of all the pixels in $\mathbb{R}_{ij}$. Var_{ij} is the variance of the circular region $\mathbb{R}_{ij}$ centred at pixel–(i,j) and with cardinality n. The result of everyone of these filters is shown in Fig. 2. Regarding contour filters, the best results where obtained with Canny,

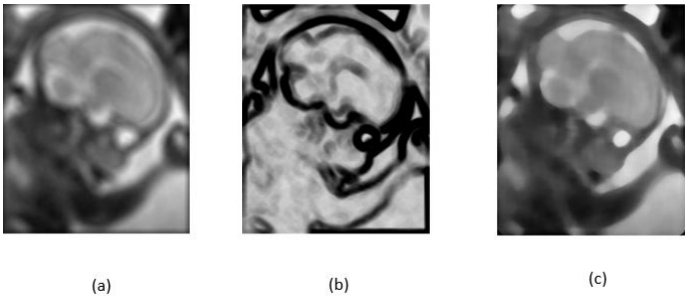


Figure 2: (a) Media filter; (b) Variance filter; (c) Median filter;

since the contours of the brain, intracranial space, and the fetal body are all clearly expressed and highlighted. The image filtered by the Canny operator is one component of the input vector of the K-nn algorithm and is also used to validate the border of the regions of interest identified by our algorithms. The other components of the input vector are the mean (or median) and variance images.

3 K-nn Algorithm

The K-nn algorithm is a non-parametric method for classification and regression whose input are the closest training examples in the feature space. The output of the algorithm when is used for classification, that is the case here, is an association to one or several classes, i.e., an object is classified by a plural vote of its neighbours and then assigned to the most common class among its closest neighbours. Neighbours are taken from a set of objects for which the class is known, the TS. The algorithm approximates the functions locally, and the entire calculation is deferred until the function is evaluated. Since this algorithm depends on the distance used in the classification task, the normalization of the training data can significantly improve its accuracy. A useful technique may be to assign weights to the neighbours, which gives a larger contribution of the closest neighbours to the average. In K-nn algorithm, K is a user-defined constant of the number of nearest neighbours. When $K = 1$ the algorithm is known as the nearest-neighbour algorithm. The training examples are vectors in a multidimensional feature space, each with a class label. The training phase of the algorithm consists only in storing the features' vectors and class labels of the training samples. Suppose we have a set of n pixels, everyone with a d-dimensional feature vector, $X_i \subset TS \in \mathbb{R}^d$, $i = 1, \cdots, n$. These are classified and partitioned into c–classes, where Y_i is the class label value of the i^{th} sample, $Y_i \in \mathrm{N}_1^c = \{1, 2, \cdots, c\}$. An unlabelled vector is classified by assigning to it the label which is most frequent among its K nearest neighbours belonging to TS. Consider some norm $\|\cdot\|$ on $\mathbb{R}^d$, e.g. Euclidean distance, Hamming distance, Manhattan distance or Minkowski distance, and a non-classified pixel of index k with features vector $X_k \subset US$ (Unlabelled Set). Hence, let a sequence of ordered pixels of TS, such that $\left\|X_{I(1)} - x_k\right\| \leq \left\|X_{I(2)} - x_k\right\| \leq \cdots \leq \left\|X_{I(K)} - x_k\right\|$, where I is the list of indices of the K nearest neighbours training pixels. The classification of point k is done by the majority vote of its K neighbours. The score value is $V_{ik} = \sum_{j=1}^{K} u_{i,I(j)} w_{I(j),k}$, where k-point belongs to the i^{th} class. u_{ir} is the membership value of the r training point that belongs to the i^{th} class (e. g. $u_{ir} = 1$ if $Y_r = i$ and $u_{ir} = 0$ otherwise). w_{jk} is the weight of every j^{th} nearest-neighbours. Finally, the class with most votes is taken as the prediction, i.e., $Y_k = i$ when $V_{ik} > V_{jk}, \forall i \neq j$. The quality of the classification of the unmarked samples X_k will be better if the result of the vote is concentrated in a certain class, by expressive majority voting. This measure is the *certainty belief factor (CBF)*: $S_{ik} = V_{ik}^{\alpha} / \sum_{j=1}^{c} V_{jk}^{\alpha}$, where α is a shape parameter. This measure can be see as a confidence value of the classification process. Values of S close to 1 reveal high level of confidence and express goodness of the classification.

4 Application of K-nn algorithms and results

Two distinct versions of K-nn algorithm are implemented. First, the segmentation of the brain image is done in a non-iterative way by a plain K-nn algorithm. In a second version (RK-nn), an iterative version of the algorithm reinforces the classification process. In every iteration, the pixels with lower CBF are rejected and not classified, and pixels with higher S-value are added to the TS in the next iteration. Also, TS pixels with lower mean weight values are discarded from the TS. Then, at the new iteration, the non-classified points are subject to a new classification process. Both algorithms were tested on the same image and training data, with $K = 50$. The components of the input vector X are the mean and variance intensity of a circular region centred in the each pixel and the value of the contour filter. For the RK-nn, the threshold value for inclusion or rejection to/from TS are, respectively, 0.8 and 0.2.The results obtained with the first version are shown in Fig. 3: brain (1), intracranial space (2) and partial fetal body (3). All pixels of the image are classified into these three classes. However, pixels that belong to the border regions or edges of structure have low levels of CBF. In the second version, the method

Figure 3: Non-iterative version: Segmentation into (1) brain; (2) intracranial space;(3) partial fetal body.

terminated at the 3rd iteration. For the specified threshold values, 50% of the pixels were well classified in first iteration, 12.1% in the second and the remaining pixels in the last one. The results are shown in Fig. 4, where white zones represent high membership value of the image region to the class. Black regions represent no pixels in the class. With this version, the confidence results of the classification process improved around 20% for the well classified pixels.The results obtained with both versions of the algorithm were compared with the K-mean clustering algorithm and other K-nn algorithms and showed better confidence results.

Figure 4: Iterative version: Segmentation into 1) brain; (2) intracranial space;(3) partial fetal body.

5 Conclusion and future work

In this work, two versions of the K-nn algorithm were used for segmentation of an fetus brain MRI. The first version is a modified version of the standard algorithm and the second is an iterative version of the first. Both algorithms produced good results to determine sub-regions of the image, although the second one presented a higher confidence value. Besides the good results already obtained, the performance of both versions of the method can be improved by taking other features of the images as input of the K-nn classifier. The presented study illustrates the capabilities of the this type of methods to support obstetricians and general practitioners to assess the fetus and, in particular, its brain.

References

[1] A. Makropoulos, S. Counsell and D. Rueckert. A review on automatic fetal and neonatal brain MRI segmentation. *NeuroImage*, 170:231–248, 2017.

[2] T. Cover and P. Hart. Nearest neighbour pattern classification. *IEEE Transactions of Information Theory*, IT-13 (1):21–27, 1967.

[3] D. Levine *et al.* Compedium of fetal MRI @ONLINE, 2002. URL `http://radnet.bidmc.harvard.edu/fetalatlas/atlas.html`.

[4] P. Habas *et al.* Atlas-based segmentation of the germinal matrix from in utero clinical mri of the fetal brain. volume 11, pages 351–8, February 2008.

[5] S. Tourbier *et al.* Automatic brain extraction in fetal mri using multi-atlas-based segmentation. In Sébastien Ourselin and Martin A. Styner, editors, *Medical Imaging 2015: Image Processing*, volume 9413, pages 248 – 254. International Society for Optics and Photonics, SPIE, 2015.

Direct Georeferencing of Fire Front Aerial Images using Iterative Ray-Tracing and a Bearings-Range Extended Kalman Filter

Bernardo Santana[2]
bernardo.santana@tecnico.ulisboa.pt
Alexandre Bernardino[1]
alex@isr.tecnico.ulisboa.pt
Ricardo Ribeiro[1]
ribeiro@isr.tecnico.ulisboa.pt

[1] Institute for Systems and Robotics
Instituto Superior Técnico
Lisbon, Portugal

[2] MSc Student,
Instituto Superior Técnico
Lisbon, Portugal

Abstract

This paper discusses the design and implementation of the Iterative Ray-Tracing algorithm for forest fire georeferencing using aerial imagery, a Global Positioning System (GPS), an Inertial Measurement Unit (IMU) and a Digital Elevation Model (DEM). Taking into account that measurement errors are amplified by the target distance, an Extended Kalman Filter (EKF) is proposed to filter multiple observations of the same object of interest. This filter extracts the bearings and range information from the geometric relation between the target and the camera in a local coordinate system. A performance comparison is done with a Cubature Kalman Filter (CKF) considering possible linearization errors induced by the EKF.

In order to validate the georeferencing and filtering algorithms, an experiment was conducted. A mobile phone was used to acquire GPS, IMU and 14 images of a target. An average position error of 74.483m was obtained at an average distance of 605m. Applying the Bearings-Range EKF and CKF reduced the error to 33.620 and 33.820, respectively.

1 Introduction

Forest fires are increasingly becoming a frequent problem in modern day society. Their destructive potential makes them a serious concern and a challenge for firefighting authorities.

Fire propagation models have already been studied that take into account weather variables such as wind [7] and also the terrain type [6]. However, these models usefulness is limited since no fire geolocation algorithm has been developed for this purpose. Henceforth, the aim of this work is to fill in this gap and develop a georeferencing algorithm based on images and telemetry recorded by an aerial vehicle. This images are assumed to be pre-processed to identify the pixels that correspond to fire.

1.1 Related Work

Forlani et al. [3] apply direct georeferencing by using the on-board Global Navigation Satellite System with the Real-Time Kinematic option with Structure from Motion and Bundle Adjustment. No ground control points are used. This methodology is, however, not suitable in a forest fire scenario, where the lack of differentiated texture and smoke prevents feature extraction and matching.

Conte et al. [2] propose an image registration approach by pattern-matching the images collected from a Micro Aerial Vehicle with satellite imagery. Multiple measurements are taken and recursive least square filter is applied. Similarly to [3], this technique relies on feature extraction, and is therefore unreliable in a forest fire environment.

Ponda et al. [8] develop a Line-of-Sight Bearings-Only EKF for target localization. This requires, however, a prior knowledge of the target's position, which is not reviewed in that work. Xu et al. [10] propose the same measurement model using a CKF instead, considering possible linearization errors induced by the standard EKF. To determine an initial approximation of the target's position, the Iterative Photogrammetry (IP) algorithm [9] is used. In spite of being efficient, the IP method can diverge when the incidence angle is smaller than the profile inclination angle.

Leira et al. [5] propose the intersection of the optic ray with a flat surface. This generalization, however, is not suitable in rough terrains, as seen in [10].

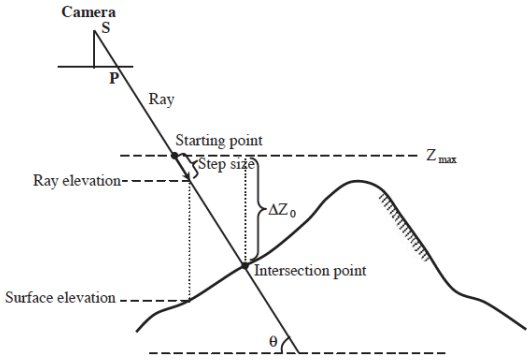


Figure 1: Iterative Ray-Tracing (adapted from [9]).

2 Georeferencing Algorithm - Iterative Ray-Tracing

The proposed georeferencing algorithm is the Iterative Ray-Tracing (IRT) [9], presented in Figure 1, and the DEM used is the EU-DEM v1.1 [1], with a spatial resolution of 25 meters and georeferenced in EPSG:3035. Since the purpose of this work is to output the geodetic coordinates of the target, this map is converted to the EPSG:4326.

The IRT works by extending the optic ray with a step size until it hits the surface. A GPS and IMU are needed to define the origin and direction of this ray, respectively, in a local NED frame. The intersection is detected when the point elevation is equal or smaller than the elevation of the DEM.

Multiple upgrades were introduced in the basic IRT, including a dynamic step size, to increase the accuracy of the algorithm. Furthermore, the starting iteration point was set as the intersection of the ray with the maximum elevation of the loaded DEM. It is expected that the aerial vehicles will operate at heights greater than the local terrain, and this can reduce the number of iterations considerably. Finally, bilinear interpolation was implemented to refine the elevation of the queried point. Ghandehari et al. [4] concluded in their work that for DEM's with finer resolutions, such has the EU-DEM v1.1, this type of interpolation achieves good results with low processing times.

3 Bearings-Range Filter

3.1 Target Dynamic Model

In this work, the target is assumed to be stationary. Therefore, its dynamics can be modeled by $\mathbf{t}_{k+1} = \Phi_{k+1|k}\mathbf{t}_k + \mathbf{Q}_k$, where $\mathbf{t}_k$ represents the target position, $\Phi_{k+1|k}$ the state transition matrix and $\mathbf{Q}_k$ the process covariance matrix:

$$\Phi_{k+1|k} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad \mathbf{Q}_k = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}. \tag{1}$$

3.2 Bearings-Range Measurement Model

The measurement model is given by $\mathbf{z}_{k+1} = \mathbf{h}(\mathbf{t}_{k+1}) + \mathbf{R}_k$, where $\mathbf{z}_{k+1}$ is the new measurement, $\mathbf{h}$ is the non-linear measurement function and $\mathbf{R}_k$ is the measurement noise covariance matrix.

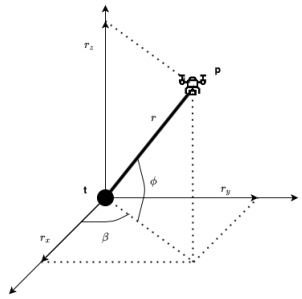


Figure 2: Bearings-Range model geometry.

$$\mathbf{h}(\mathbf{t}_{k+1}) = \begin{bmatrix} \beta \\ \phi \\ r \end{bmatrix} = \begin{bmatrix} \tan^{-1}\left(\frac{p_y - t_y}{p_x - t_x}\right) \\ \tan^{-1}\left(\frac{p_z - t_z}{\sqrt{(p_x-t_x)^2+(p_y-t_y)^2}}\right) \\ \sqrt{(p_x-t_x)^2+(p_y-t_y)^2+(p_z-t_z)^2} \end{bmatrix}, \quad (2)$$

where β and ϕ are the azimuth and elevation angles, respectively, and r is the distance between the target, $\mathbf{t}$, and the aerial vehicle, $\mathbf{p}$, as presented in Figure 2.

4 Experiment

The unavailability of telemetry and imagery data from an aerial vehicle led to the development of an alternative methodology to validate the proposed algorithm. A mobile phone was used to record GPS, IMU and image data along a pedestrian path. The natural elevation of *Serra dos Candeeiros*, near *Porto de Mós*, *Leiria*, was used to capture images of a target at a lower height, so as to simulate the overview of an aerial vehicle. A total of 14 images were acquired at an average target distance of 605 meters. For the filtering, the IRT result for the first observation is used to initialize the filter state, $\mathbf{t}_0$. The initial state covariance $\mathbf{P}_0$ and measurement noise covariance matrix $\mathbf{R}_k$ were tuned to

$$\mathbf{P}_0 = \begin{bmatrix} 20^2 & 0 & 0 \\ 0 & 50^2 & 0 \\ 0 & 0 & 1^2 \end{bmatrix}, \quad \mathbf{R}_k = \begin{bmatrix} 5^2 & 0 & 0 \\ 0 & 5^2 & 0 \\ 0 & 0 & 10^2 \end{bmatrix}. \quad (3)$$

Details on the EKF and CKF algorithms can be found in [8] and [10], respectively.

The position error is defined as $\mathbf{e_p} = \mathbf{t} - \hat{\mathbf{t}}$, where $\hat{\mathbf{t}}$ is the estimated target. σ_x, σ_y and σ_z are defined as the square root of the filter state covariance matrix diagonal. The results of the standalone IRT, EKF and CKF are summarized in Table 1.

Method	$\|\|e_p\|\|$ [m]	$\|\|\sigma_{x,y,z}\|\|$ [m]
IRT	74.483	n.d.
IRT+EKF	33.620	7.2497
IRT+CKF	33.820	7.2502

Table 1: Norm of the average position error for the standalone IRT and for the final correction of the EKF and CKF.

The IRT results presented in Figure 3 evidence a bias along the positive East direction, which then influences the estimated positions of the EKF and CKF.

5 Conclusions

In this paper, the IRT is proposed as a georeferencing algorithm using the EU-DEM v1.1. Expecting measurement errors from the GPS and IMU, a bearings-range filtering algorithm was developed, with a performance comparison between the EKF and CKF. Preliminary results using the data collected with a mobile phone show evidence of bias susceptibility. This may be due to the non-ideal preliminary experimental setup using a line of sight more parallel to the ground when compared to the more vertical one from an aerial vehicle. Furthermore, the 14 images were captured at approximate positions, limiting the new information added to the filtering algorithm. Still, an improvement of 41 meters is achieved on the 74 meter

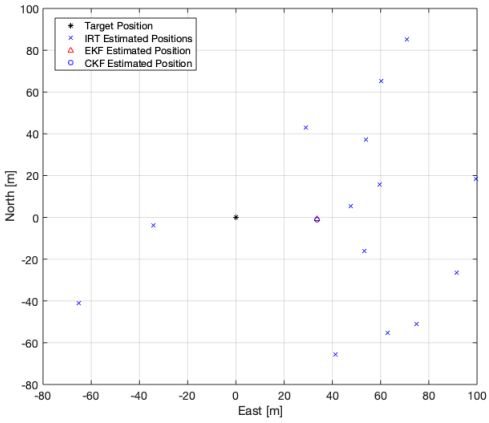


Figure 3: Real and estimated target positions by the IRT, EKF and CKF algorithms.

average position error of the standalone IRT. There is no clear advantage in using the CKF over the EKF for this measurement model.

Acknowledgements

This work was supported by FCT with the LARSyS - FCT Project UIDB/50009/2020 and project FIREFRONT (PCIF/SSI/0096/2017).

References

[1] EU-DEM v1.1. URL `https://land.copernicus.eu/imagery-in-situ/eu-dem/eu-dem-v1.1`. visited on 26-07-2020.

[2] Gianpaolo Conte, Maria Hempel, Piotr Rudol, David Lundström, Simone Duranti, Mariusz Wzorek, and Patrick Doherty. High accuracy ground target geo-location using autonomous micro aerial vehicle platforms. *AIAA Guidance, Navigation and Control Conference and Exhibit*, pages 1–14, 2008.

[3] Gianfranco Forlani, Fabrizio Diotri, Umberto Morra di Cella, and Riccardo Roncella. Indirect UAV Strip Georeferencing by On-Board GNSS Data under Poor Satellite Coverage. *Remote Sensing*, 11(15):1765, 2019.

[4] Mehran Ghandehari, Barbara P Buttenfield, and Carson J Q Farmer. Comparing the accuracy of estimated terrain elevations across spatial resolution. *International Journal of Remote Sensing*, 40(13): 5025–5049, 7 2019.

[5] Frederik S. Leira, Kenan Trnka, Thor I. Fossen, and Tor Arne Johansen. A ligth-weight thermal camera payload with georeferencing capabilities for small fixed-wing UAVs. *2015 International Conference on Unmanned Aircraft Systems, ICUAS 2015*, pages 485–494, 2015.

[6] A. M.G. Lopes, A. C.M. Sousa, and D. X. Viegas. Numerical simulation of turbulent flow and fire propagation in complex topography. *Numerical Heat Transfer; Part A: Applications*, 27(2):229–253, 2 1995.

[7] A. M.G. Lopes, L. M. Ribeiro, D. X. Viegas, and J. R. Raposo. Simulation of forest fire spread using a two-way coupling algorithm and its application to a real wildfire. *Journal of Wind Engineering and Industrial Aerodynamics*, 193(July):103967, 2019.

[8] Sameera S. Ponda, Richard M. Kolacinski, and Emilio Frazzoli. Trajectory optimization for target localization using small unmanned aerial vehicles. *AIAA Guidance, Navigation, and Control Conference and Exhibit*, (August), 2009.

[9] Yongwei Sheng. Comparative evaluation of iterative and non-iterative methods to ground coordinate determination from single aerial images. *Computers and Geosciences*, 30(3):267–279, 2004.

[10] Cheng Xu, Daqing Huang, and Jianye Liu. Target location of unmanned aerial vehicles based on the electro-optical stabilization and tracking platform. *Measurement*, 147, 12 2019.

Computational Analysis of Nonverbal Communication Cues in Group Settings

Rui Frazão[1]
ruifilipefrazao@ua.pt
Samuel Silva[12]
sss@ua.pt
Sandra Soares[3]
sandra.soares@ua.pt
António J. R. Neves[12]
an@ua.pt

[1] DETI
University of Aveiro, Aveiro, Portugal
[2] IEETA
University of Aveiro, Aveiro, Portugal
[3] CINTESIS.UA
WJCR.UA
Dept. of Education and Psychology
University of Aveiro, Aveiro, Portugal

Abstract

Human communication is a major field of study in psychology and social sciences. Topics such as emergent leadership and group dynamics are commonly studied cases when referring to groups. Group settings experiments are usually studied in conversational and collaborative tasks environments in order to study the communication process in small groups. Former study methods involved human analysis and manual annotation of other's behaviors in communication settings. Later studies try to replace time consuming and failure prone annotations by resorting to computational methods.

For that purpose, we propose a multimodal approach capable of using a broad range of nonverbal communication in a complementary way in order to allow the quantification of nonverbal aspects from video data. This paper presents a framework capable of contributing to a direct increase in human knowledge about the human communication process, involving data transformation processes in order to transform raw feature data into humanly understandable meanings.

1 Introduction

Communication is a natural, omnipresent process in human lives. Since birth, humans use signals, sounds, movements and expressions to communicate with others. Human communication is defined as the process of human being's interaction to other's behaviors. When one thinks about human communication, verbal communication is what comes to mind first. This type of communication is done verbally and relies on the use of words and phrases to convey meanings. However, information passed during the communication process lies not only on what the sender and receiver are transmitting verbally but also through their behavior.

For a long period of time, the human communication subject has been studied mainly by the psychology and social sciences areas, manually classifying behaviors and annotating datasets without computational methods involved in the process. Nowadays, there are computational methods capable of extracting features from human communication situations, thus allowing a deeper analysis of the communication process.

The research reported in this paper aims to contribute to the study and understanding of the human communication phenomenon. The main objective is to create a framework grounded on a critic analysis of current literature, on how to develop a computational system capable of quantifying nonverbal aspects from video data, following a multimodal approach, by analyzing different nonverbal features simultaneously in a complementary way, and thus broadening the analysis of the communication process context, contributing to richer information, as the message is eventually only understood in full when all its parts are considered.

2 Background

Researchers have defined nonverbal communication by identifying characteristics that constitute it [4]. The set of signals transmitted via a particular medium or channel is called "Code". The various codes in combination form the structure of nonverbal communication as known today. This codes are often defined by the human sense or senses they stimulate and/or the carrier of the signal [2]. Table 1 enumerates the different types of codes along with some of the features they represent.

Code Type	Code Name	Features
Visual	*Kinesics*	Facial expressions; Head movements; Eye behavior; Gestures; Posture; Gait
Auditory	*Vocalics* or *Paralinguistics*	Dialect; Pitch; Tempo; Dysfluencies; Intonation
Body	*Attractiveness*	Appearance; Adornments; Olfatics
Contact	*Proxemics*, *Haptics*	Space; Distance; Touch
Time	*Chronemics*	Timing
Place	*Artifacts*	Environment Objects

Table 1: Types of nonverbal communication codes and corresponding code names and constituent features.

As the data needed to analyze human communication is multimodal by nature, following a multimodal approach tends to achieve better results compared to single modalities [7], yet, the processing of multidimensional data also constitutes a problem due to the need of large computational power and optimized algorithms.

Multimodal studies can follow either a complementary or redundant analysis. Complementary approaches focus on broadening the analysis of information emitted from multiple communication channels, while redundant tend to seek validation to an assumption or conclusion, using different information from various communication features.

3 Computational Approach to the Study of Nonverbal Communication

For a specific use case, a multidisciplinary research group, having members from engineering and psychology backgrounds collected a custom dataset from an experiment conducted by the Department of Psychology of the University of Aveiro regarding the influence of conflict in group collaborative tasks. As such, the experiment consists in two tasks, where only in the latter, the conflict is induced.

The input data is composed of three camera sensors pointed to a table where four subjects sit and perform LEGO construction tasks. To limit the image processing to the relevant image regions reducing computational times and unwanted artifacts, the image data is clipped to a custom-defined region of interest, and created a method to match subjects in different perspectives.

Considering the nonverbal codes and its features and the available raw pose and facial features extracted from the video data, it is possible to discard the chance of including auditory and time features, as the former requires audio data and the latter is not relevant as each task must be done within a specific time frame, which is monitored by the experiment staff and thus would not provide any additional information.

There are two main elements to extract: Pose and Facial Landmarks. There are several state-of-the-art methods for human pose and facial keypoints extraction. We use OpenPose[3] and DensePose[5] for pose extraction and OpenFace[8] for facial landmarks extraction.

After the feature extraction phase, it is be possible to match nonverbal cues to behaviors and understand the meaning of those behaviors. By identifying relevant features, an individual analysis of every element's features is done, followed by the combination of each of their corresponding feature vectors in the following step.

Having extracted the feature vectors and correspondent meanings, it was necessary to study how to quantify the nonverbal cues. Determining metrics that apply to the specific use case is also of value to the task and can be calculated by transforming and/or combining the extracted vectors. This is a major step in understanding the correlation between behaviors and the human communication process. Some of this information can also

be presented overlaid on the original image data in order to better analyze and understand the behavior during the communication phenomenon.

Visual features

Visual features are mainly linked to posture and positioning of the involved subjects. Such features involve: Facial features, which allow the analysis of head movement and direction and emotion related data, and can be directly related to the level of interest of a subject in a determined task [7, 8]. In order to quantify emotion, a naive approach based on combining the activation of facial muscles was followed [9]; Body expansiveness, which is usually correlated to the perception of influence, power and dominance, can be quantified as the occupied area both horizontally and vertically by the polygon involving the furthest pose keypoint; Group activity, as it is intended to understand how is group energy affected by each experiment condition and if, as a consequence of the existence of conflict in a group, habits tend to vary and how as this can be an indicator of subjects' disengagement. As a way of quantifying the activity, two approaches were taken: Motion Energy Image as proposed by [1, 7], and analysis of the keypoint movement between frames.

Body features

Body features are not easy to quantify based on nonverbal behavior. Studies show strong correlation between physical attractiveness, body and face symmetry, and social and cognitive attributes as the most relevant characteristics in the attractiveness field [4, 6]. Not being able to retrieve data covering those aspects, only antropometric information could possibly be used. Although it is shown in literature that there is a correlation between some physical attributes and, perceived competence and leadership in group settings [4], this information alone is not considered relevant to this specific case.

Contact features

Contact and visual codes can be considered highly correlated as features such as occupied space, distance and touch are measured taking posture and gestures into account. Proximity-related features such as overlap and distance between group subjects (intragroup distance) can also be correlated as the closer the group is, the bigger is the overlap. Overlap is calculated based on the subjects' occupied areas intersections, and intragroup distance is given by the distance between subjects. These features are considered representative of group cohesiveness and consequently, the level of engagement in the task. Cohesiveness is an important matter in study of group dynamics, as it is positive involvement behavior [4, 10].

Place features

Artifact related aspects can easily be extracted in this particular use case, as the experiment features the handling of objects displayed in the experiment environment. Such interactions can provide enriched information about the subjects' behavior and engagement in the group tasks. In this specific use case, the interaction with the displayed objects is measured by the subjects' distance to the center of the table.

4 Results and Conclusion

The most important output of the work described in this document is a framework capable of quantifying the described nonverbal aspects in a group setting from video data, with the intent to aid field professionals analyzing a group's behavior, offering overlaid visual information on top of the original data, and generating plots of the quantified features. The dataset used in this work was fully annotated by this framework.

Figure 1, on top, presents the visualization tool developed to display the processed feature data regarding nonverbal communication aspects on top of the original input data, contributing to a easier analysis of such aspects. Here is demonstrated both the overlaid information of subject's keypoints and overlap. At the bottom, a plot illustrating the comparison between groups' intragroup distance along the experiment.

Some aspects that would improve the results obtained in the work would be the use of higher quality image data and the use 3D information in order to be possible to analyze other types of features that are not possible to quantify in a 2D space, such as body orientation.

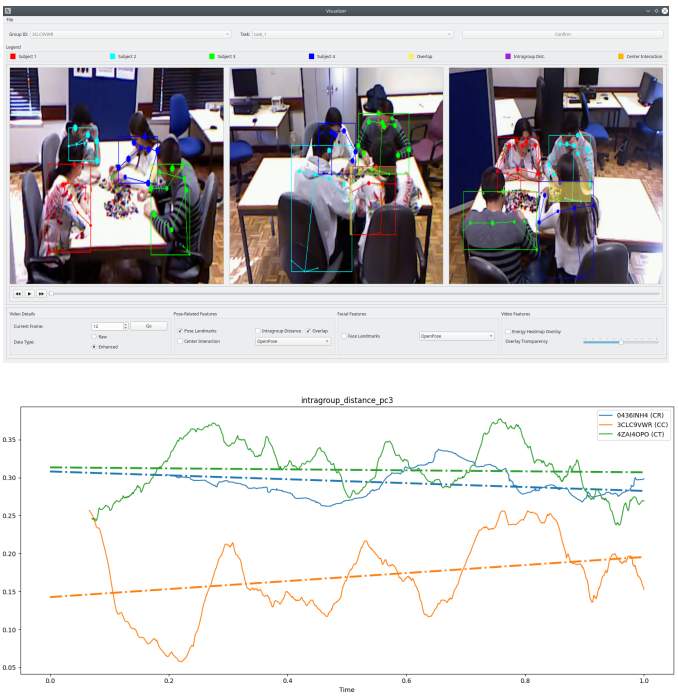


Figure 1: Top: Visualization tool displaying keypoint position and overlap features; Bottom: Comparison between groups intragroup distance along the experiment.

References

[1] Aaron F. Bobick and James W. Davis. The recognition of human movement using temporal templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(3):257–267, 2001. ISSN 01628828. doi: 10.1109/34.910878.

[2] Judee K Burgoon, Laura K. Guerrero, Kory Floyd, Laura K. Guerrero, and Kory Floyd. *Nonverbal Communication*. Routledge, jan 2016. ISBN 9781315663425. doi: 10.4324/9781315663425.

[3] Zhe Cao, Tomas Simon, Shih En Wei, and Yaser Sheikh. OpenPose: Realtime multi-person 2D pose estimation using part affinity fields. *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, pages 1302–1310, 2018. doi: 10.1109/CVPR.2017.143.

[4] Laura K. Guerrero. Interpersonal functions of nonverbal communication. In *Interpersonal Communication*, pages 53–75. DE GRUYTER, Berlin, Boston, 2014. ISBN 9783110276794. doi: 10.1515/9783110276794.53.

[5] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. DensePose: Dense Human Pose Estimation In The Wild. feb 2018.

[6] J. L. Rennels. *Physical attractiveness stereotyping*, volume 2. Elsevier Inc., 2012. ISBN 9780123849250. doi: 10.1016/B978-0-12-384925-0.00099-7.

[7] Dairazalia Sanchez-Cortes, Oya Aran, Marianne Schmid Mast, and Daniel Gatica-Perez. A nonverbal behavior approach to identify emergent leaders in small groups. *IEEE Transactions on Multimedia*, 14:816–832, 2012. ISSN 15209210. doi: 10.1109/TMM.2011.2181941.

[8] Tadas Baltrusaitis. *Automatic facial expression analysis*. PhD thesis, University of Cambridge, 2005.

[9] Sudha Velusamy, Hariprasad Kannan, Balasubramanian Anand, Anshul Sharma, and Bilva Navathe. A method to infer emotions from facial Action Units. In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2028–2031. IEEE, 2011. doi: 10.1109/ICASSP.2011.5946910.

[10] Bin Zhu, Xin Guo, Kenneth E. Barner, and Charles Boncelet. Automatic group cohesiveness detection with multi-modal features. In *ICMI 2019 - Proceedings of the 2019 International Conference on Multimodal Interaction*, pages 577–581. Association for Computing Machinery, Inc, oct 2019. doi: 10.1145/3340555.3355716.

Learning to Grasp Objects in Virtual Environments through Imitation

Alexandre Filipe
Thesis student,
Instituto Superior Técnico

Alexandre Bernardino, Plinio Moreno
IST Robotics Department,
Instituto Superior Técnico

Abstract

To improve the quality of robot grasps we propose the use of human demonstrations as a guide for the robot to follow, a process often referred to as imitation learning. To do so a human subject would perform grasps and manipulation tasks on a virtual environment, using a glove with sensors capable of capturing the entirety of the hand motions. Our goal is to develop a predictive model, which uses past and current joints' information to estimate the forthcoming joints' positions. Our model is a Recurrent Neural Network that generates the joint positions for a virtual robot, replicating the demonstrated task. To ensure the objects are well grasped, the task is segmented in two phases, after and before the object is considered grasped, avoiding the model continuing a task with an object not well grabbed, dropping it or not even lifting as a result.
Using this model, trained with the recorded demonstrations, we guide the virtual robot to perform a series of simple manipulation tasks, and manage to do so with an good success rate.
Also, to allow anyone intending to try and test their own imitation algorithms, we will provide our virtual environment and complete dataset of demonstrations freely.

1 Introduction

Even with today's standards of what robots can achieve, robotic manipulation still stands as a huge challenge, due to the great number of degrees of freedom present in a human hand, leading to the community using a claw/gripper or a 3-finger hand instead [1][2].

To train a robot to grasp and manipulate an object, several approaches have been used, from physics based to trial and error methods, but a group that has shown some promise are imitation based methods, where a human subject demonstrates the grasping task and, using some form of machine learning, the robot tries to replicate the same task.

To record said demonstrations, again, several methods are employed, from depth-image [1] to video [2][3], but we propose the use of a virtual environment, as some works before did [3], as it is easier to capture data and does not suffer the common issues of having objects, including the hand, obstructing the view of the camera. To capture the hand movements, in case the recording isn't made through video, a range of tools are used, including motion controllers included with VR headsets [1] and keyboard/console controllers [2] but, to fully capture the complexity of the human hand, we are going to use gloves with sensors [4][5].

2 Process

With our objective of teaching a robot through demonstrations, being these demonstrations performed in a virtual environment using a glove with sensors, we must first have a virtual environment capable of capturing these. To do so we used Unity3D, a well-known game engine, to create a simple virtual environment that consists on a table with several objects to interact with (Figure 1). The physics interactions would be processed by Unity3D's own physics engine.

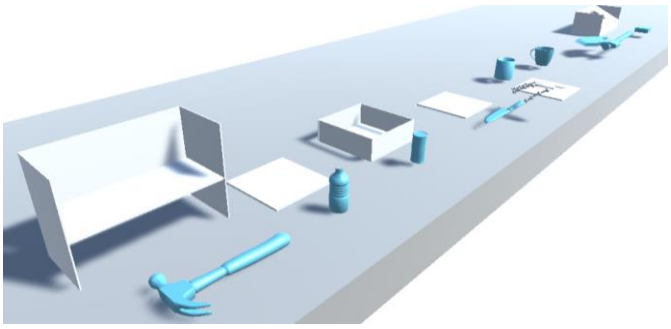


Figure 2 - Virtual Environment

To interact with this virtual environment we used the VMG 35 Haptic glove (Figure 2), a glove containing sensors to measure how much each finger joint is bent and 2 gyroscopes, one on the hand and another on the wrist, to read the rotational position of the hand.

Figure 1 - VMG 35 Haptic glove

With everything set it was now possible to record the demonstrations. Such is done by recording the hand and object positions every fifth of a second, creating a sequence of stages/iterations that represent the demonstration, and then writing this data to a txt file. For each task, having each one consisting of grasping an object and, in some, interacting with a second object, 200 demonstrations were recorded. To create some variance the initial position of the objects is randomized at the start of each recording, inside a plausible area.

Having a considerable set of demonstrations, it is now needed to use this to allow a robot to use their information to perform the task. To achieve this we used an LSTM [6], a recurrent neural network that had already been shown to be successful in robot manipulation [2]. To be precise, we found it better to use 2 LSTMs, segmenting the demonstration in 2 parts, before the object is considered to be grabbed (the reaching segment), and after (the manipulation segment), having each LSTM training a single segment. This was done as a single LSTM would sometimes lead to the hand almost grabbing the object and them proceeding with the task without the object in hand. With 2 LSTMs the hand will only advance to the manipulation segment after the grab condition is met, continuing to try to grab the hand until then.

The input of the LSTM consists of the current state along with the previous 9 states. The state consists of hand position, object position and hand's limbs angular positions. The output of the LSTM is the prediction of the next state. By performing this prediction in a recurrent way, the model will be able to recreate the training tasks.

The data format consists, as mentioned before, of a sequence of iterations, being each iteration represented by 20 values that contain the values of each finger joint bend and the object and hand position. To streamline and minimize the number of input values for the LSTMs we record the relative position (spatial and angular), on the reaching segment the relative position of the hand to the object and on the manipulation segment the relative position of the hand (with the object grabbed) to another object we will interact with, or a generic position, in case no second object is used.

To estimate the forthcoming joint position, we use a kind of a regression predictive model, where the LSTM is trained with the mean squared error loss.

To make the trained LSTMs guide the virtual hand to perform task first we choose a starting position, from which the reaching LSTM will recurrently output the following iteration, having the virtual hand in Unity3D following these values. The instant the object is considered to be grabbed (this condition is verified if the thumb and at least one opposable finger is in contact with the object), a flag is sent from Unity3D informing that from that point on we will continue the reproduction using the manipulation LSTM, continuing this one predicting the following iterations recurrently.

3 Simulation Results

To test the quality of our method we trained and tested for 4 tasks: grabbing a bottle and placing it on a base, grabbing a can and placing it sideways on a box, grabbing and bringing a mug to a base and grabbing and placing a hammer on a shelf.

At the start of each test the hand and the object to be grabbed starts at a random positions (inside plausible values). Afterwards the LSTMs will try to predict the following iterations, using the process mentioned before, and try to perform the task they were trained for. If the objective of the task is achieved (that is, if the hammer is placed on the shelf, for example), then the reproduction is considered a success.

After 20 trials of testing for each task the success rate was as follows:

Table 1 - Test results

Bottle	85%
Can	80%
Mug	75%
Hammer	75%

These results show promise for this method, achieving the common values for this kind of work.

4 Next Steps

Having the model capable of guiding the virtual hand we think it would also be of interest to try to export the model to guide a real life robot to perform the tasks that has been demonstrated on the virtual environment.

To anyone that wants to test their own manipulation training methods our virtual environment and our complete dataset of demonstrations can be found in github.com/alexamor/thesis.

Acknowledgements

This work was supported by FCT with the LARSyS - FCT Project UIDB/50009/2020.

References

[1] T. Zhang, Z. McCarthy, O. Jow, D. Lee, X. Chen, K. Goldberg, and P. Abbeel, “Deep imitation learning for complex manipulation tasks from virtual reality teleoperation,” in2018 IEEE InternationalConference on Robotics and Automation (ICRA), May 2018, pp. 5628–5635

[2] R. Rahmatizadeh, P. Abolghasemi, A. Behal, and L. Boloni, “Learning real manipulation tasks from virtual demonstrations using lstm,” arXiv preprint arXiv:1603.03833, 2016.

[3] Y. Liu, A. Gupta, P. Abbeel, and S. Levine, “Imitation from observation: Learning to imitate behaviors from raw video via context translation,” in 2018 IEEE International Conference on Robotics and Automation (ICRA), May 2018, pp. 1118–1125.

[4] A. Rajeswaran, V. Kumar, A. Gupta, J. Schulman, E. Todorov, and S. Levine, “Learning complex dexterous manipulation with deep reinforcement learning and demonstrations,” inRobotics Proceedings - Robotics: Science and Systems XV, 2017.

[5] Sundaram, S., Kellnhofer, P., Li, Y. et al. Learning the signatures of the human grasp using a scalable tactile glove. Nature 569, 698–702 (2019). https://doi.org/10.1038/s41586-019-1234-z

[6] Felix A. Gers, Nicol N. Schraudolph, and Jurgen Schmidhuber. Learning precise timing with LSTM recurrent networks. Journal of Machine Learning Research (JMLR), 3(1):115–143, 2002

Fire and Smoke Detection using CNNs trained with Fully Supervised methods and Search by Quad-Tree

Gonçalo Perrolas
goncalo.perrolas@tecnico.ulisboa.pt
Alexandre Bernardino
alex@isr.tecnico.ulisboa.pt
Ricardo Ribeiro
ribeiro@isr.tecnico.ulisboa.pt

Instituto de Sistemas e Robótica
Instituto Superior Técnico
Lisboa, Portugal

Abstract

Wildfires prevail as one of the most destructive and uncontrollable natural disasters for man-kind. The fire-fighting combat team can greatly benefit from reliable information about the current position of active burning areas. It is also important to detect as early as possible the fire ignition sources through the smoke columns produced.

In this work we use aerial images taken from drones of the wildfires to detect fire and smoke using deep neural networks. A set of tests is presented to evaluate the performance of the network used in the fire and smoke detection. To solve the multi-scale detection problem we use a Quad-Tree method in the search task.

The proposed system shows an adequate performance in real drone aerial images.

1 Introduction

This article presents initial results of the Firefront Project (`www.firefront.pt`). Its main objective is to create a support system to the fire combat teams. The system will transmit valuable information about the wildfire in real-time to the ground teams using aerial vehicles to capture images of the scene. Next, the images taken, are used to detect fire and smoke using a convolution neural network able to segment the respective areas.

It comes as a challenge the detection of variable size portions of both fire and smoke, and due to the small input size image that can be feed to the common CNNs an extra algorithm was used for doing detection on smaller image sections dynamically using a Quad-Tree search method.

1.1 State-of-the-Art

The problem of fire and smoke detection using RGB images has been highly studied on the last few years, using different approaches and techniques. Existing methods can be divided into two big groups: classic methods and methods that use deep learning.

The classic methods tend to use the RGB components of images via histogram analysis, to evaluate what colours are more related to fire areas. One example is the method used by Cruz *et al*. [1] that relies on indexes based on RGB components to boost fire detection and then applies a threshold to binarize the image into the classes. This process is very time and computing power efficient but results in poor detection performances and this method is very constrained to the environment conditions of the images taken and the camera characteristics. The active system CICLOPE that is operating in Portugal, developed by Batista *et al*. [2] uses a background subtraction method using fixed cameras to do smoke detection. This technique makes use of the advantages of comparing consecutive frames with the same background, showing a good performance but, as we intend to use aerial vehicles to gather the images, a different method must be used.

At the other side of the spectre, deep neural networks can be used in this detection problem, as this area of investigation is evolving increasingly year after year. The most common type of neural network used in these types of problems is convolutional neural networks. These are able to extract image information on an abstract level allowing to choose the characteristics that better represent fire or smoke portions of the image. Frizzi *et al*. [3] used a type of CNN using weak supervision training to do fire and smoke segmentation. Training is a crucial stage while working with CNNs and requires a large set of images paired with the corresponding labels for fire and smoke. The results shown proved that the trained network was over-fitted to the data because of the small dataset used, which is an important challenge due to the lack of fire and smoke datasets available. There are a lot of CNN extensions, one of them being the R-CNN. Barmpoutis *et al*. [4] made use of this type of CNN and got good results, it showed an efficiency way of detecting one instance of fire with a bounding box. On that work it was used an interesting and complete fire dataset called Corsican Fire Dataset [5] which was pretty useful for the training of our networks.

2 Methodology

Firstly the system requires an aerial vehicle to capture the images from a wildfire or the beginning of one. Then, two independent systems are used, one in charge of fire segmentation and other for smoke segmentation. As input to these systems we are going to fed small portions of the images (patches) to be able to detect small or big areas of fire/smoke on those images. The process of slicing the image into smaller patches it is going to be done using a Quad-Tree methodology. An overall diagram of the complete system and the corresponding results are shown in Figure 1.

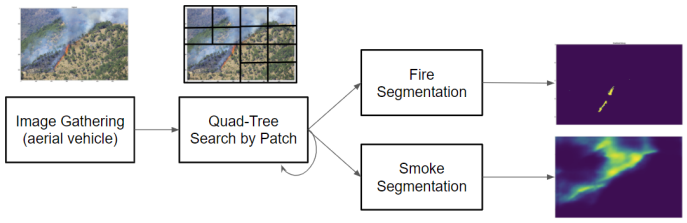


Figure 1: Global Diagram of the system.

2.1 Search by Quad-Tree

Quad-Tree methods are often used to partition a two-dimensional space by recursively subdividing it into four quadrants or regions. This is useful for doing a dynamic search for fire or smoke occurrences on the images, starting first with larger scale patches and, if nothing is detected, then moving on to smaller patches by slicing the previous ones in four segments. By using this method we solve the issues related to the multi-scale nature of fire/smoke, areas, in large patches are difficult to be detected by the neural networks (small input size) and patches that are almost filled completely with fire or smoke can have dubious detection results due to the lack of more external features that define the phenomenon.

2.2 Patch Processing (Classification + Segmentation)

The detection system is composed of two different neural networks, one for classification and other for segmentation. The classification network chosen was a SqueezeNet model [6] due to its level of accuracy compared to state-of-art models like Alex-Net while having a lot fewer parameters making the model more suitable for smaller datasets. For segmentation we used U-net [7], one of the most used state-of-the-art models for semantic segmentation.

By using a classification stage before doing segmentation the overall performance is increased due to the fact that it is easier to create a more complete dataset with image level labels then with pixel level. This reduced the number of false detection on the segmentation.

The overall logic of the system can be explained as following: each patch of the image produced by the Quad-Tree stage is given as input to

the SqueezetNet [6] to determine if the patch contains fire or smoke. If a phenomenon is detected then the patch moves along to the segmentation stage where the areas of fire/smoke are segmented by the U-Net [7]. If nothing is detected in the classification stage the patch is not segmented and the process moves to the next patch in the sequence.

2.3 Datasets

Both networks need a set of images with the corresponding labels identifying the respective classes. For the classification network the labels needed are on image level (the images does/doesn't contain fire/smoke) for the segmentation model we need pixel level labels. The images gathered for the fire dataset mainly came from three different sources: Corsican Dataset [5] (RGB images with pixel wise labelling), smaller datasets found online and a batch of images gathered online that were manually labelled to extend as much as possible the size of the dataset. The smoke dataset consists also of datasets with pixel wise labels found online and some more images segmented manually.

The datasets for classification include the ones used for the segmentation training together with some more images, as this one doesn't need extensive labelling. In table 1 we present a small overview of the datasets used for the training phase where is identified the corresponding number of images.

Classification	Fire	Positive	800 imgs
		Negative	520 imgs
	Smoke	Positive	500 imgs
		Negative	300 imgs
Segmentation	Fire	Containing Fire	700 imgs
		Negative	450 imgs
	Smoke	Containg Smoke	300 imgs
		Negative	60 imgs

Table 1: Overview of the Dataset

Good negative cases for fire are sunrises, sky with red tonalities, red objects and rooftops. For smoke the main concern was related to clouds due to the very challenging similarities.

3 Results

The dataset was divided as three different subsets randomly, one for training, one for validation and one for testing purposes (70%,20%,10%). After training all the four networks, we evaluated the performance of the methods here proposed. On table 2, the full system segmentation results are shown. The results show a good general performance, although it is rather important to refer that the classes fire/negative and smoke/negative are unbalanced within each image, typically the images have a smaller area of fire or smoke. We can also conclude that the smoke detection performance is a bit worse than the fire. In terms of processing time, for an average size image (800 x 500), the processing takes roughly 4 secs, running on a Tesla K80 GPU and Intel Xeon CPU.

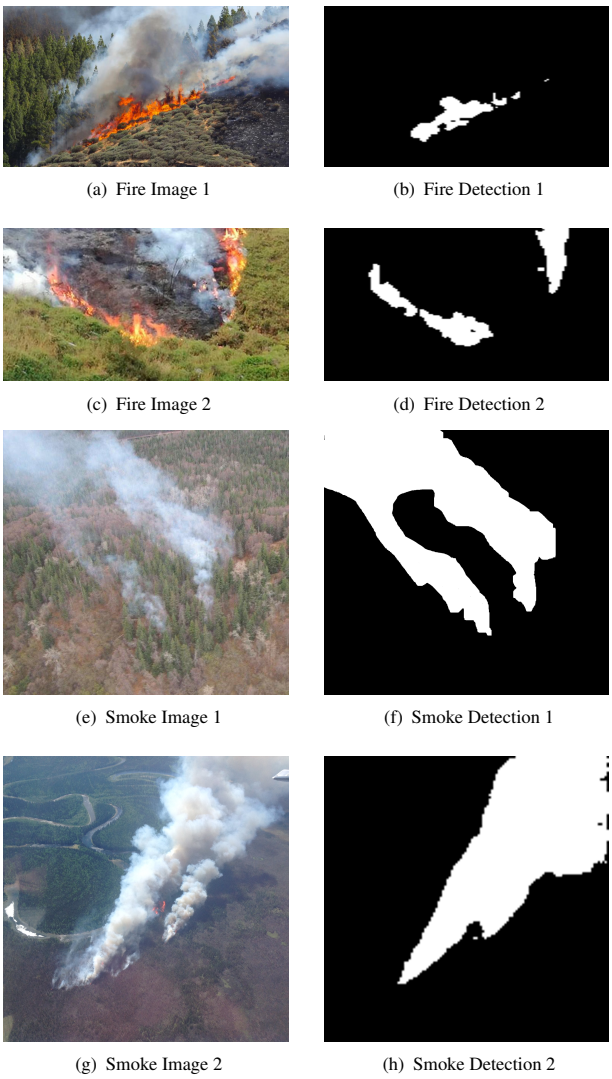
	Avg. IoU	Pixel Accu.
Fire	0.8692	0.8348
Smoke	0.8404	0.7519

Table 2: Performance Metrics on test set

On figure 2, are shown some examples of the results of the detections produced by the developed system.

4 Conclusions

The datasets are the core asset of the neural networks used in this approach. As we improved our dataset we observed a boost on the robustness of the performance of the detection for both fire and smoke. It would be a huge progress in the development of this area of study if it was created a centralised and accessible database with more diverse set of images. The results shown here prove that the system is reliable and can be applied to real life monitoring of fire situations and it has potential to represent a big improvement in the way we deal with fire.



(a) Fire Image 1 (b) Fire Detection 1

(c) Fire Image 2 (d) Fire Detection 2

(e) Smoke Image 1 (f) Smoke Detection 1

(g) Smoke Image 2 (h) Smoke Detection 2

Figure 2: Examples of results produced for both fire and smoke detection

Acknowledgements

This work was supported by FCT with the LARSyS - FCT Project UIDB/50009/2020 and project FIREFRONT (PCIF/SSI/0096/2017)

References

[1] H. Cruz M. Eckert J. Meneses J. Martínez. Efficient forest fire detection index for application in unmanned aerial systems (uass). *Sensors (Basel)*, June 2016.

[2] M. Batista B.Oliveira P.Chaves J. Ferreira T. Brandão. Improved real-time wildfire detection using a surveillance system. In *Proc. of the World Congress on Engineering 2019*, 2019.

[3] S. Frizzi R. Kaabi M. Bouchouicha J. Ginoux E. Moreau F. Fnaiech. Convolutional neural network for video fire and smoke detection. *IECON 2016 - 42nd Annual Conference of the IEEE Industrial Electronics Society, Florence*, 2016.

[4] P. Barmpoutis K. Dimitropoulos K. Kaza and N. Grammalidis. Fire detection from images using faster r-cnn and multidimensional texture analysis. *ICASSP - IEEE*, 2019.

[5] T. Toulouse L. Rossi A. Campana T. Celik M. Akhloufi. Computer vision for wildfire research: an evolving image dataset for processing and analysis. *Fire Safety Journal*, 2017.

[6] Forrest N. Iandola Song Han Matthew W. Moskewicz Khalid Ashraf William J. Dally Kurt Keutzer. Squeezenet: Alexnet-level accuracy with 50x fewer parameters and <0.5mb model size. 2016.

[7] O. Ronneberger P. Fischer T. Brox. U-net: Convolutional networks for biomedical image segmentation. 2015.

Assessment of Motor Compensation Patterns in Stroke Rehabilitation Exercises

Ana Rita Cóias
ana.coias@tecnico.ulisboa.pt
Alexandre Bernardino
alex@isr.tecnico.ulisboa.pt

Institute for Systems and Robotics
Instituto Superior Técnico, ULisboa
Lisboa, PT

Abstract

The increasing demand concerning stroke rehabilitation and in-home exercise promotion requires objective methods to assess patients' quality of movement, allowing progress tracking and promoting consensus among treatment regimens. In this work, we propose a method to detect diverse compensation patterns during exercise performance with 2D pose data to automate rehabilitation programs monitorization in any device with a 2D camera, such as tablets, smartphones, or robotic assistants.

1 Introduction

With the escalating demands towards stroke rehabilitation and the increase of in-home exercise recommendations [2], the need for new means to evaluate patients' motor performance has risen [4, 7]. In conventional assessment tests, therapists assess movement quality based on observation, thus being highly subjective [4]; with the degree of experience implying distinct treatment approaches [7]. Quantitative and objective methods allow patients' progress tracking, impaired movements' understanding, and formulation of standard therapy regimens [4, 6].

Patients' physically impaired often exhibit compensation behaviors to accomplish a task. Motor compensation is the presence of new movement patterns derived from the adaptation or substitution of old ones, which might help patients' execute a task [5]. New patterns can include the use and activation of additional or new body joints and muscles. Most typical compensation behaviors are trunk displacements, rotation, and shoulder elevation. These functional strategies are commonly observed in reaching and are highly related to severe impairment levels [5].

Early on the recovery process, the use of compensation strategies promotes patients' upper limb participation in task performance. However, their persistence may obstruct real motor function recovery and must be reduced during therapy through appropriate exercise instructions [5].

In this work, we present a method to assess quantitatively motor compensation from video frames during upper limb exercise performance. We have created a labelset (Table 2) for each video frame of the dataset regarding the observed compensation patterns. We then explore two methods to assess these patterns based on 2D pose data enabling this kind of analysis with widely available RGB cameras.

2 Related Work

When conceiving quantitative methods to assess movement quality, researchers carry out the kinematic study of 3D pose data to track patients' progress, enhance in-home therapy, and bring consensus among therapists' evaluation. Kinematics delineates body movements over space and time, giving information on linear and angular displacements. Prior works usually explore joint angular motion and trunk displacements. Some studies determined which kinematic variables better describe motor impairment and identify upper limb disability levels through PCA analysis [6]. Others assessed the quality of the upper extremity movement with machine learning methods [4, 7]. However, existing methods do not detect distinct compensation patterns and are based on 3D pose data, which limits its wide applicability in off-the shelf computational devices.

3 Learning to Assess Motor Compensation

Considering stroke survivors with one weakened side of the body, we assess motor compensation through individuals' body parts' 2D pose data extracted from video frames. To accomplish this task, we execute the following steps: body keypoints extraction and selection, data normalization, and multilabel classification to determine the compensation patterns observed among the video frames. We present a rule-based (RB) classification method, which works as our baseline approach and a Neural Network (NN) that assesses compensation through the body keypoints.

3.1 Feature Extraction and Selection

To extract the body joints' 2D pose data, we use the OpenPose [1], a software library that provides the location of 25 body keypoints in the image coordinate system. Each keypoint is denoted by $p_j^t = [x\ y]'$, where j denotes a body joint and t the frame number.

We consider two scenarios (S1 and S2) concerning patients' position in front of the camera: one facing the recording camera (S1) and the other with the patient's affected arm facing the camera (S2). According to [4], we select the joints shown in Figure 1 to describe patients' movements, which are held by the RB and NN methods. The head keypoints, $j \in [15, 18]$, are held for the RB method, in addition to the selected joints, to overcome the lack of 3D data by head size variation. Considering a multi-person setting (with the patient under evaluation and a caregiver), we select the patient assuming he/she is the closest person to the center of the image.

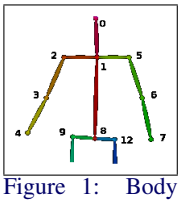


Figure 1: Body keypoints.

3.2 Data Normalization

In a real-world setting, subjects have body parts' of different sizes and are not placed at the same place regarding the camera. For this reason, we normalize the keypoints. First, we apply rigid body transformation from the image coordinate system, $\{I\}$, to the body coordinate system, $\{B\}$, in which the patient's joint 8 is the origin. This step considers the patients' affected side. For S1, the ${}^BX$ axis is directed to the affected side. For S2, the ${}^BX$ axis is directed to the patients' front. Additionally, we normalize each resultant keypoint coordinates to the spine length measured in $t = 1$. For the NN, to give the non-affected side as a reference, we mirror the joints to the ${}^BX$ axis positive side, aligning both sides. For RB, each keypoint moves regarding other specified keypoint.

3.3 Kinematic Variables

We compute kinematic variables for the RB approach to describe motion patterns similar to [4, 6]. However, as we work with 2D positional data, we do not have information about patients' movements in depth. This way, we formulate hypotheses to detect the different compensation patterns: trunk moving forward, trunk rotation, shoulder elevation, and other trunk displacements, such as trunk tilt and trunk moving backward. More specifically, for both scenarios S1 and S2 - the formulated hypotheses and respective kinematic variables are summarized as follows.

Trunk Forward/Backward: S1 - observed changes in patient's head size, ΔH^t (H^t - head area in $t > 1$); S2 - spine angular and linear displacements, $a^t(p_8^1, p_1^1, p_1^t)$ (a^t - angle between three joints) and $d_x^t(p_1^t, p_1^1)$ (d_x^t - displacement in X).

Trunk Rotation: S1 - simultaneous angular displacements of both shoulders, $a^t(p_2^1, p_1^1, p_2^t)$ and $a^t(p_5^1, p_1^1, p_5^t)$; S2 - absolute changes in the observed chest length, $|\Delta d^t(p_2^t, p_5^t)|$ (d^t - Euclidean distance between two joints) or[1] shoulder displacement regarding joint 1 in X, $d_x^t(p_{2/5}^t, p_1^t)$.

Shoulder Elevation: S1 - shoulder elevation angle $a^t(p_{2/5}^1, p_1^1, p_{2/5}^t)$; S2 - shoulder displacement regarding joint 1 in Y, $d_y^t(p_{2/5}^t, p_1^t)$ (d_y^t - displacement in Y).

Trunk Tilt: S1 - spine angular displacement $a^t(p_8^1, p_1^1, p_1^t)$; S2 - absolute changes in patient's head size, $|\Delta H^t|$.

[1] In S2 patients can show their chest or be completely aside

	Exercise	Scenario	P_{min}
$E1$	*'Bring a Cup to the Mouth'*	S1	83.83%
$E2$	*'Switch a Light On'*	S1	91.4%
$E3$	*'Move a Cane Forward'*	S2	98.15%

Table 1: The three exercises and percentage of single labeled frames.

Label	$IRLbl_{E1/E2/E3}$	Label	$IRLbl_{E1/E2/E3}$
'0: Trunk Forward'	-/-/3.54	*'3: Other'*	4.93/5.55/-
'1: Trunk Rotation'	16.23/19.25/-	*'4: Normal'*	1/1/1
'2: Shoulder Elevation'	2.15/3.03/15.77	-	-

Table 2: Considered labels and *IRLbl* metric for each one.

3.4 Classification Approaches

While exercising, a stroke survivor can describe multiple compensation moves. Thus, we consider this problem a multilabel classification problem and learn the different compensation patterns observed in a video frame. We explore two approaches: a RB method and a NN that learns the observed patterns based on the keypoints position.

The former method is a set of *if-then* rules (e.g. *'2'* if shoulder angle is above a threshold) applied to the kinematic variables and ending in the class labels [3, 8]. The latter is an ensemble of two classifiers seizing to respect label dependency and overcome label imbalance [8]. The first classifier (C1) executes binary classification, verifying compensation existence. If there is compensation, the second classifier (C2) performs multilabel classification to determine the pattern. Here we apply binary relevance *One-vs-Rest*, which considers each label independently. Afterward, we join the classification results into the multilabel output.

4 Method Validation

To validate our method, we use the rehabilitation exercise videos from Lee *et al*. work [4]. We validate the formulated hypotheses to assess compensation through the kinematic analysis and present the classification results with our baseline classifier. To validate the NN ensemble, we apply *Leave-One-Subject-Out* (LOSO) cross-validation (CV).

4.1 The Multilabel Dataset

The dataset consists of videos with 15 stroke survivors performing an average of 10 movement trials of three upper extremity exercises (E1, E2, and E3), detailed in Table 1. We assigned to every video frame multiple labels (Table 2). Label *'3'* includes trunk tilt and moving backward. Label *'4'* holds movements with no compensation or the resting state. As shown in Table 1, the dataset is almost single labeled - high percentage of single labeled frames, P_{min}. Regarding label imbalance, in Table 2, the *IRLbl* metric shows the ratio between the occurrences of the most frequent label and each label. We can see that, for the three exercises, label *'4'* is the most frequent, $IRLbl = 1$. For E1 and E2, *'1'* is poorly represented, $IRLbl \gg 1$, with only one patient exhibiting this compensation pattern. For E3, the less representative label is *'2'*.

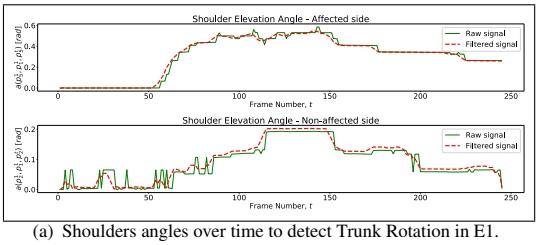
4.2 Kinematic Variables

We validate the hypotheses formulated to assess compensation from 2D positional data. Figures 2(a) and 2(b) show the variation over time of three kinematic variables used to assess compensation behaviors, without 3D data. In Figure 2.(b) trunk rotation is assessed in E1 with the simultaneous angular displacement of both shoulders. In Figure 2(b) shoulder elevation is detected in E3 through shoulder displacement in Y regarding joint 1.

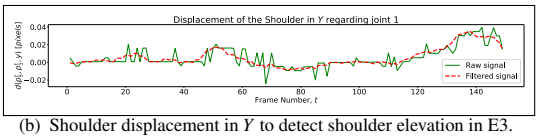
4.3 Classification Results

When applying the RB and performing LOSO CV to the NN approach, we obtained the average results given in Table 3. For the NN we explored one to two layers with 16, 64, and 96 hidden units with adaptive learning rate. We apply *'ReLu'* for C1 and *'Tanh'* for C2 activation functions and *'Adam'* optimizer with mini-batch size of 5.

As we can see in Table 3, the NN method performs better for the E2 and E3, with a higher P_{min} value, meaning the RB handles better E1. This suggests that the NN may work better in single labeled cases. Also, the high levels of standard deviation in both methods suggest that the approaches could benefit from more exercise examples from more patients to improve generalization ability.



(a) Shoulders angles over time to detect Trunk Rotation in E1.



(b) Shoulder displacement in Y to detect shoulder elevation in E3.

Figure 2: Examples of kinematic variables variation over time.

	Precision	Recall	$F1-score$	HammingLoss
$E1_{RB}$	**0.756 ± 0.14**	**0.783 ± 0.12**	**0.767 ± 0.12**	**0.11 ± 0.06**
$E2_{RB}$	0.555 ± 0.17	0.666 ± 0.17	0.602 ± 0.17	0.187 ± 0.08
$E3_{RB}$	0.697 ± 0.27	0.71 ± 0.26	0.701 ± 0.26	**0.126 ± 0.11**
$E1_{NN}$	0.692 ± 0.23	0.678 ± 0.25	0.679 ± 0.24	0.187 ± 0.15
$E2_{NN}$	**0.673 ± 0.21**	**0.675 ± 0.19**	**0.668 ± 0.19**	**0.182 ± 0.11**
$E3_{NN}$	**0.785 ± 0.22**	**0.783 ± 0.21**	**0.783 ± 0.22**	0.153 ± 0.14

Table 3: Average results for the rule-based (RB) and Neural Network (NN) methods.

5 Conclusions

We conclude that our method assesses distinct compensation patterns during upper extremity exercise performance pretty well from 2D pose data. In future work we want to leverage more data to achieve better label distribution and representativeness.

Acknowledgements

This work was supported by FCT with the LARSyS - FCT Project UIDB/50009/2020 and project IntelligentCare – Intelligent Multimorbidity Management System (Reference LISBOA-01-0247-FEDER-045948), co-financed by the ERDF – European Regional Development Fund through the Lisbon Portugal Regional Operational Program – LISBOA 2020 and by the Portuguese Foundation for Science and Technology – FCT under CMU Portugal.

References

[1] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[2] T. M. Damush, L. Plue, T. Bakas, A. Schmid, and L. S. Williams. Barriers and facilitators to exercise among stroke survivors. *Rehabilitation Nursing*, 32(6), 2007.

[3] M. Kubat. *An Introduction to Machine Learning*. Springer, 2017.

[4] M. H. Lee, D. P. Siewiorek, A. Smailagic, A. Bernadino, and S. Bermúdez I Badia. Learning to assess the quality of stroke rehabilitation exercises. In *International Conference on Intelligent User Interfaces, Proceedings IUI*, pages 218–228. Association for Computing Machinery, 2019.

[5] M. F. Levin, J. A. Kleim, and S. L. Wolf. What do motor "recovery" and "compensation" mean in patients following stroke? *Neurorehabilitation and Neural Repair*, 23:313–319, 2009.

[6] M. A. Murphy, C. Willén, and K. S. Sunnerhagen. Kinematic variables quantifying upper-extremity performance after stroke during reaching and drinking from a glass. *Neurorehabilitation and Neural Repair*, 25(1):71–80, 2011.

[7] E. V. Olesh, S. Yakovenko, and V. Gritsenko. Automated assessment of upper extremity movement impairment due to stroke. *PLoS ONE*, 9(8), 2014.

[8] G. Tsoumakas and I. Katakis. *Multi-Label Classification*, volume 3. Springer, 2011.

Exploring the Impact of Color Space in 6D Object Pose Estimation

Nuno Pereira
nuno.pereira@ubi.pt
Luís A. Alexandre
luis.alexandre@ubi.pt

Departamento de Informática
Universidade da Beira Interior
NOVA LINCS
6201-001, Covilhã, Portugal

Abstract

6D pose estimation is an open challenge due to complex world objects and many possible problems when capturing data from the real world, *e.g.*, occlusions, and truncations. Getting the best input data to the deep learning methods is critical, for example, light can alter the features that these methods extract from the objects. Not obtaining accurate poses of the objects can lead to poor experiences in augmented reality scenarios or can lead to a fail grasping task of a robot. To try to avoid these issues, we investigate the impact of color spaces in 6D object pose estimation. For that, we evaluated RGB, Grayscale, HSV, and the HSV individual channels to study which color space would perform better in the 6D pose estimation task. We increased the accuracy of a method in 7.11% by using the HSV color space instead of the frequently used RGB.

1 Introduction

In computer vision, the ambient light can be a notable problem. It can create artifacts, alter the colors or cause shadows in the captured scene therefore constituting a problem in many computer vision algorithms.

The RGB color space is widely used, although it does not represent the color as humans perceive it. If we want to isolate an object just using color in the image, it is hard to do in RGB because there may be many similar colors in the image.

The HSV color space has three channels similar to RGB but instead of Red, Green, and Blue we have Hue, Saturation and Value, or intensity. The Hue channel represents the color. For example, red is a color but light red or dark red is not. The saturation channel is the amount of color present. It differentiates the pale red from the pure red. Finally, the value or intensity represents the brightness of the color, light red or dark red. So in the Hue channel, each color has its own value the entire red is a particular value. The lightness or darkness of the color does not affect the hue channel, so this channel is useful to extract specific colors from images. In real photographs, you will obtain varied saturation throughout the images depending on the intensity of the color present in them. The intensity channel shows the brightness of the colors and this channel usually has much influence by the light source.

With the required automation needed to help humans in production lines, robots need to work in a collaborative mode and work in non-restricted environments so the capacity to understand the scene and the objects within is becoming a must. The most common task that robots do is object grasping, which is a task that has been tackled by many researchers because it needs to be as fast as possible and precise so there is no damage in the objects. Performing gasping in a non-restricted and cluttered environment, *e.g.*, bin picking, is a complex problem to tackle.

6D pose estimation is a task in computer vision that detects the 6D pose (3 degrees of freedom for the position and the other 3 for orientation) of an object. A 6D pose is as important in robotic tasks as in augmented reality, where the pose of real objects can affect the interpretation of the scene and the pose of virtual objects can also improve the augmented reality experience. It can also be useful in human-robot interaction tasks like learning from demonstration and human-robot collaboration.

Estimating the object's 6D pose is a challenging problem due to the diversity of objects that exist and how they appear in the real world. Obtaining the data to retrieve the 6D pose is a problem, as RGB-D data can be hard to obtain for certain types of object, *e.g.*, fully metallic objects, and meshed office garbage bins. Another problem during the data capture is the light present in the scene because it can generate noise or reflection in the objects. Distinct light sources can make deep learning methods extract different features from the same object this being a problem because if we want that the method learns these types of light sources we need to capture data from each type of light source. More specific, 6D pose estimation requires massive amounts of data to have good performance in real-world applications and even when the methods are trained in these big datasets they usually tend to fail or have a greater error in the real world because the most common and well-structured datasets have well-controlled light sources.

Figure 1: Qualitative results on the LineMOD Dataset. These results were obtained using MaskedFusion with HSV color space as input data. The red dots represent the object keypoints of the estimated 6D pose projected onto the RGB image.

To prevent this situation, we propose an alternative approach to this type of method by analyzing other color spaces. Color spaces like HSV are uncommon in the 6D pose estimation area of research. We test if using other color spaces in the most common dataset LineMOD [1] will increase the performance of 6D pose estimation methods while not increasing significantly its training or inference times.

2 Methodology

We use MaskedFusion [4] as a framework for 6D pose estimation in our experiments. It is one of the best-performing methods in the state-of-the-art.

MaskedFusion consists of three sub-tasks that executed sequentially estimates the 6D pose of an object presented in the scene. Initially, it uses a semantic segmentation method to detect and generate masks for each object presented in the scene. Then for each object segmented it crops the RGB image, depth image and mask. To eliminate the background around the object, a bit-wise and operation is made between the images and the mask. These segmented images are fed to a fully convolution neural network so it can regress the 6D pose of that object. After the preliminary pose is estimated, it is possible to utilize another method to refine the pose of the object. The method used in MaskedFusion is a neural network that enables it to be executed in real-time instead of other methods that are resource-heavy.

This work was supported by NOVA LINCS (UIDB/04516/2020) with the financial support of FCT-Fundação para a Ciência e a Tecnologia, through national funds, and partially supported by project 026653 (POCI-01-0247-FEDER-026653) INDTECH 4.0 – New technologies for smart manufacturing, cofinanced by the Portugal 2020 Program (PT 2020), Compete 2020 Program and the European Union through the European Regional Development Fund (ERDF)

Table 1: Results presented in this table were obtained through the training of MaskedFusion with its weights initialized as random. Italic names represent the symmetric objects. Bold values are the higher values in each line.

Objects	RGB	Grayscale	HSV	H	S	V
ape	74.29	86.67	**97.14**	67.62	34.29	82.86
bench vi.	99.03	**100.00**	99.03	88.35	89.32	99.03
camera	96.08	**98.04**	**98.04**	87.25	75.49	97.06
can	94.06	97.03	**98.02**	80.20	91.09	93.07
cat	**97.00**	95.00	**97.00**	81.00	86.00	**97.00**
driller	96.00	95.00	**99.00**	91.00	88.00	94.00
duck	62.26	93.40	**96.23**	51.89	35.85	88.68
eggbox	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**
glue	**100.00**	99.03	**100.00**	**100.00**	**100.00**	**100.00**
hole p.	91.43	95.24	**99.05**	89.52	74.29	**99.05**
iron	75.26	97.94	95.88	94.85	91.75	**98.97**
lamp	**100.00**	**100.00**	98.08	95.19	97.12	**100.00**
phone	**100.00**	99.04	**100.00**	93.27	94.23	99.04
Average	91.17	96.63	**98.28**	86.08	81.14	96.03

In our experiments, we did not use the first sub-task of the MaskedFusion. Our primary goal is to report the impact of the different color spaces and/or channels in the 6D pose estimation. Since MaskedFusion is a modular framework, it was effortless to remove the semantic segmentation sub-task and use the ground truth masks to make the operations for the crop and background removal.

To perform our tests, we choose to compare the HSV color space and each of its channels with the RGB color space. We tested MaskedFusion using the RGB, HSV, Grayscale, H (Hue), S (Saturation), and V (Value). We evaluated the MaskedFusion method in two independent roundups. In the first series of tests executed we trained the method from scratch, this means, the neural network presented in the method started with random weights, and we trained it for 150 epochs. In the second series of tests, we trained the method with RGB for 350 epochs. Furthermore, we saved the best performing weights in the validation set and use these weights to start fine-tuning the neural network for the other color channels. We fine-tuned the neural network for 150 epochs.

In our tests, we use the LineMOD [1] dataset because it is widely utilized in this area of research. It consists of 13 objects in over 18000 real images with the ground truth pose annotated. These images were captured with a Kinect camera that automatically aligns the RGB and depth images.

As in previous works in 6D pose estimation [2], [3], [5], [6], [4] we use the same evaluation metrics for the LineMOD dataset. The Average Distance of Model Points (ADD) [1] is used for non-symmetric objects and for the egg-box and glue the Average Closest Point Distance (ADD-S) [6] is used.

$$\text{ADD} = \frac{1}{m} \sum_{x \in M} \left\| (Rx + t) - (\hat{R}x + \hat{t}) \right\| \tag{1}$$

In the ADD metric (equation 1), assuming the ground truth rotation R and translation t and the estimated rotation $\tilde{R}$ and translation $\tilde{t}$, the average distance calculates the mean of the pairwise distances between the 3D model points of the ground truth pose and the estimated pose. In equation (1) and (2) M represents the set of 3D model points and m is the number of points.

For the symmetric objects (egg-box and glue), the matching between points is ambiguous for some poses. So for these cases, the ADD-S metric is used:

$$\text{ADD-S} = \frac{1}{m} \sum_{x_1 \in M} \min_{x_2 \in M} \left\| (Rx_1 + t) - (\hat{R}x_2 + \hat{t}) \right\| \tag{2}$$

3 Results

In Table 1 and 2, we present the results of MaskedFusion in the LineMOD test set. These results were calculated using the ADD (equation 1) and ADD-S (equation 2) metric. In Table 1, we present the results where the MaskedFusion neural network was trained for 150 epochs with weights initialized as random values.

In Table 1, its shown that the best performing color space is the HSV, as it performed higher on average. Specially for the first object, it achieved less error overall. HSV color space also achieved the best accuracy in 10 out of 13 objects.

Table 2: Results presented in this table were obtained by fine-tuning. Italic names represent the symmetric objects. Bold values are the higher values in each line.

Objects	RGB	Grayscale	HSV	H	S	V
ape	88.57	90.48	96.19	92.38	**97.14**	94.29
bench vi.	**99.03**	97.09	**99.03**	97.09	**99.03**	**99.03**
camera	**99.02**	**99.02**	**99.02**	98.04	95.10	**99.02**
can	96.04	**98.02**	97.03	**98.02**	93.07	96.04
cat	99.00	99.00	**100.00**	99.00	**100.00**	99.00
driller	96.00	95.00	92.00	**98.00**	**98.00**	94.00
duck	95.28	91.51	94.34	**96.23**	93.40	91.51
eggbox	99.06	**100.00**	**100.00**	99.06	**100.00**	**100.00**
glue	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**
hole p.	99.05	**100.00**	**100.00**	**100.00**	98.10	98.10
iron	97.94	96.91	**100.00**	97.94	93.81	95.88
lamp	99.04	99.04	99.04	**100.00**	98.08	**100.00**
phone	94.23	94.23	97.12	**98.08**	97.12	94.23
Average	97.08	96.93	**97.98**	**97.98**	97.16	97.01

In Table 2, we present the results for the executed tests with fine-tuning. The results presented were obtained by using the best-performed weights in the evaluation set during 350 epochs and then we used these weights to fine-tune the MaskedFusion for the other color spaces. Fine-tuning took 150 epochs and then we evaluate the method in the test set. On average the HSV color space and the Hue color channel had the lowest average error in the LineMOD dataset. Both of these colors had seven objects in which they performed higher than the other color spaces/channels.

During inference, we took an average 0.014 seconds to estimate the 6D pose of an object. Our experiments took on average more 0.002 seconds to estimate the 6D pose of an object comparing its execution time with the RGB color space that did not need any color space conversion. These times were obtained using a computer with SSD NVME, 64GB of RAM, an NVIDIA GeForce GTX 1080 Ti and Intel Core i7-7700K CPU.

4 Conclusion

Sometimes using different color spaces aid in specific computer vision tasks. In these evaluations we discovered that using the HSV color space can help MaskedFusion achieve less error overall, if the same number of training epochs are used as when training using RGB images.

Training MaskedFusion for 150 epochs from the random weights in the RGB color space we achieved on average 91.17% accuracy and using the same setup but only changing to HSV color space we achieved 98.28%, a substantial improvement.

These tests might even achieve better results when dealing with real-world scenarios, since, the LineMOD dataset, as others, used a controlled light environment during the capture of the data creating the best possible scenario for each image presented in it. We suspect that the advantage of HSV over RGB can be even greater when pose estimation is performed in uncontrolled environments, and this will be a topic for future work.

References

[1] Stefan Hinterstoisser, Stefan Holzer, Cedric Cagniart, Slobodan Ilic, Kurt Konolige, Nassir Navab, and Vincent Lepetit. Multimodal templates for real-time detection of texture-less objects in heavily cluttered scenes. In *ICCV*, 2011.

[2] Wadim Kehl, Fabian Manhardt, Federico Tombari, Slobodan Ilic, and Nassir Navab. Ssd-6d: Making rgb-based 3d detection and 6d pose estimation great again. In *ICCV*, 2017.

[3] Sida Peng, Yuan Liu, Qixing Huang, Xiaowei Zhou, and Hujun Bao. Pvnet: Pixel-wise voting network for 6dof pose estimation. In *CVPR*, 2019.

[4] Nuno Pereira and Luís A. Alexandre. MaskedFusion: Mask-based 6d object pose estimation. In *19th IEEE International Conference on Machine Learning and Applications (ICMLA 2020)*, December 2020.

[5] Chen Wang, Danfei Xu, Roberto Zhu, and Lu. Densefusion: 6d object pose estimation by iterative dense fusion. In *CVPR*, 2019.

[6] Yu Xiang, Tanner Schmidt, Venkatraman Narayanan, and Dieter Fox. Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes. *arXiv:1711.00199*, 2017.

Fire and Smoke Detection in Aerial Images

Bernardo Amaral[1]
bernardo.m.amaral@tecnico.ulisboa.pt
Alexandre Bernardino[1]
alex@isr.tecnico.ulisboa.pt
Catarina Barata[1]
ana.c.fidalgo.barata@tecnico.ulisboa.pt

[1] Institute for Systems and Robotics, ISR
Instituto Superior Tecnico, IST
Universidade de Lisboa,
Lisbon, PT

Abstract

Fire and smoke detection in images using image processing and deep learning techniques has proven to be a topic of high interest. Recent methods for object recognition and localization using deep learning achieve very reliable results. Methods that rely on large amounts of data with the respective annotations, which are both expensive and often subjective. To overcome this limitation, we study and implement a method to detect fire and smoke zones using only weakly supervised methods. We train a convolutional neural network based model (CNN) for object classification that relies only on image-level labels and still can learn to predict the location of fire and smoke. We demonstrate that the model is able to localize the discriminative image regions of fire and smoke despite not being trained for them.

1 Introduction

Forest fires are a scourge that every year destroy thousands of hectares of forest around the world. Forest fires have a series of effects on both the burned area and the underlying areas. The consequences go beyond the visible effects on nature and society such as the destruction of material assets and the effect on vegetation. The entire ecosystem is threatened, from fauna and flora to loss of biodiversity, soil degradation and erosion, to CO^2 emissions. For this reason it is extremely urgent to take measures to mitigate these dangers and reduce the risk of forest fires. According to [1], in Portugal on the last three years (2017-2019) about 630 thousand hectares were burnt due to forest fires.

A possible approach to create a wiser firefighting and minimize this threat is through the use of manned or unmanned aerial vehicles that collect real-time visual information from the fire site. Then the information can be feed to automatic systems that are able to locate regions of fire and smoke. This poses a considerable challenge, since neither fire nor smoke have a well-defined shape and a constant color.

The usual methods of locating/identifying objects using deep learning are trained on a large amount of fully annotated data, which means that in each image of the dataset there must be an annotation of where each class is present in the image, for example through the use of bounding boxes or pixel-level masking. Though, the creation of such annotations is very expensive and, especially in case of fire and smoke, it can be very subjective, since there are no rigid limits.

To overcome this situation we propose to use a weakly supervised method where the only needed annotation is at the image-level. Which means that each image in the dataset has only the information if each class is present in the image or not, there is no information of its location. With this approach, we can train a CNN model to classify the presence or absence of fire and smoke and still predict their correspondent localization.

2 Related work

The work done on fire and smoke detection based on computer vision presented a wide variety of methods. The big majority of them are based on color, motion, spatial and temporal features. This characteristics are very specific for fire compared to other objects. Most of them follow a common pipeline, first find moving pixels using background subtraction and then apply a color model to find fire color regions. The approach is to create a mathematical based model, defining a sub-space on a color space that represents all the fire-colored pixels in the image [2]. On this line, *Wang et al*. [3] proposed a method based on a Gaussian model learned in the YCbCr color space. *Uğur et al*. [4] added to the base pipeline a wavelet-based model of fire's frequency signature, with the idea that flames flicker with a characteristic frequency. Also, *Chine et al*. [5] added texture analysis to create *Bowfire* and prevent false positives resulted from the single use of color models. Methods using motion tracking are limited since they only work properly with fixed cameras, in surveillance scenarios [6]. They are therefore not compatible for use with aerial vehicles.

These methods depend heavily on the features delimited by the authors, which may make them too specific for a certain purpose. On the other hand, methods with deep learning, automatically learn which features are best for the given problem. This is why deep learning methods outperform them. On [7] the authors do a comparative analysis between color-model based methods versus deep learning methods. They use a logistic regression model, which is a very simple deep learning method and yet it is the one that obtains the best overall performance compared to all colour-based models. They also prove the robustness of these methods for colour changes and the presence of smoke.

Thus, considering the superiority of the deep learning methods compared to those based on colour, several authors presented methods using CNN to detect fire and/or smoke [8]. Still following the idea of using surveillance cameras, *K. Muhammad* [9] trained a SqueezeNet model for fire detection, localization, and semantic understanding of the scene of the fire. *Q Zhang et al*. [10] trained a Faster R-CNN to detect smoke in wildland forest fire by creating synthetic images with the addition of synthetic smoke to normal forest images. Also *Q. Zhang et al*. [11] propose a method to detect and localize fire using a CNN by using image patch division.

The lack of a good public accessible dataset for fire and smoke makes it hard to develop a good deep learning technique. For this reason some of the previous methods that were trained with small datasets, and even thought they used fully annotated images, might lack some robustness. In the method we propose, we can use the few datasets available [12] and complement them with as many images as we want without much effort to label them. Then we can build a robust and reliable method to detect and locate fire and smoke.

3 Methodology

Our approach is based on the powerful ability to locate objects from the convolutional layers in a CNN, using only a set of images that are annotated at image-level. We chosen a *VGG19* as our base model. We then removed the fully connected layers because they destroy the spatial integrity kept in the convolutional layers and added a *Global Average Pooling* (GAP) and *Sigmoid* layers. The *Sigmoid* layer makes the model behave as a common one-vs-all classifier for each class. Therefore, for each image we will always have the prediction if there is fire, smoke, both at the same time or none at all. This is very important since fire and smoke are highly correlated. *Figure 1* summarizes our proposal.

To obtain the location, we applied the methodology proposed in [13] to our specific case. The idea is to create a Class Activation Map (CAM) that can be used to localize the network's attention on the input images for fire and smoke even though the networks have only been trained on image-level labels. This is done by gathering the information from the features maps on the last convolutional layer. Thus, we had a *GAP* layer before the last convolutional layer so that we can weight the importance of each feature map for the predicted class(es). Then, we do a weighted sum of those feature maps according to the predicted class to produce the CAM. Hence, $H_c(x,y) = \sum_{i=1}^{n} w_i^c f_i(x,y)$ where H represents the CAM with the predicted location and w_i^c is the weight of the activation of the i^{th}

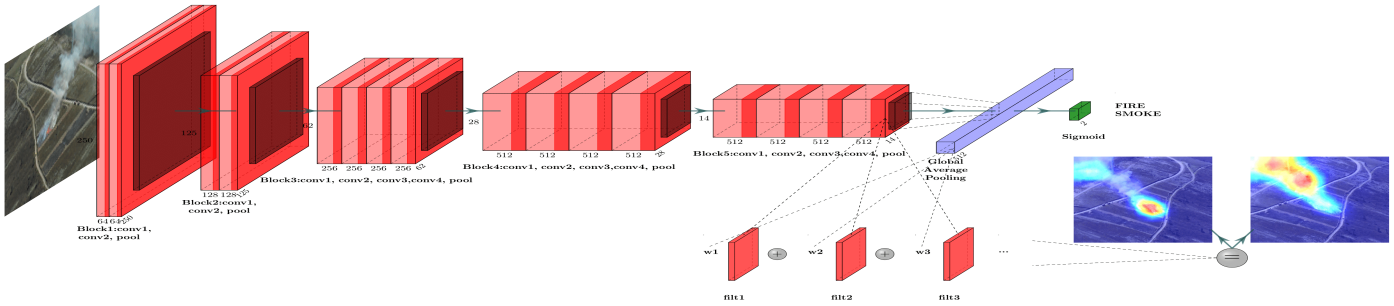


Figure 1: Global architecture

feature map $f(x,y)$ for the predicted class c.

By doing so, CAM behaves as a heatmap highlighting the areas in the image where it is more probable to be fire and/or smoke.

4 Experimental setup

To train the network, we build our own data set. As a starting point the dataset of [12] was used since it contains good examples of forest fires as well as controlled fire frames. In order to complement it, we added images from aerial views, images with only fire, images with only smoke and images of negative cases (sunsets, very orange scenes, clouds, etc.). Gathered from the web and from the Firefront project [14]. In the end, the data set is made up of 1770 images, of which 80% are for training, 12.5% are for validation and 7.5% are for testing.

Our model is exclusively trained for classification purposes and as it behaves as a one-vs-all classifier we used a binary cross-entropy loss. At any stage of the model training we use any location related loss.

5 Results

To evaluate our method, we tested our approach in terms of classification and segmentation.

For classification, we tested on our test set. The metrics presented are at the image level, the network predicts the presence of fire and smoke in the whole image. We achieved an accuracy of 92% for fire and 91% for smoke, which is a good result taking into account that we never said to the network what exactly is fire or smoke.

Regarding the segmentation, we tested only for the case of fire, using the images and their ground truth from [12]. To do this, we had to transform the CAMs into binary masks. The CAM behave like a heatmap with values between 0 and 1 in which areas with values close to 1 are more likely to belong to the class and vice versa. Thus, we have created a threshold where any value above is considered to belong to the class. Several values were tested taking into account the mean Intersection-Over-Union (IoU) . In the end we obtained a value of 0.35 for the threshold and a corresponding IoU of 0.575. To note that this a very good value taking into account that the ground truth masks are very detailed while our method, although accurate, is not so precise, is more rounded. If we apply a small dilation on the ground truth masks, we can do a more equitable comparison and achieve a mean IoU of 0.61.

6 Conclusions

Our method provides a solution for developing location capabilities when there is a lack of available data, in the specific case applied to fire and smoke. We have shown that even with only image level labels, we have been able to build a normal classification network to predict the location and have saved a lot of time on labeling.

We are aware that the heatmaps produced are not as sharp as an output of a segmentation network but are accurate in terms of location. Consequently, in a future work we could sharpen the heatmaps with the use of different available method's, for example Conditional Random Field (CRF).

Acknowledgments

This work was supported by FCT with the LARSyS - FCT Project UIDB/50009/2020 and project FIREFRONT (PCIF/SSI/0096/2017).

References

[1] PORDATA. Portugal - environment, energy and territory - territory and environmental protection - rural fires and burnt area, Accessed 10/08/2020. www.pordata.pt/en/Portugal/Rural+fires+and+burnt+area+-+Mainland-1192.

[2] Celik Turgay and Hasan Demirel. Fire detection in video sequences using a generic color model. *Fire safety journal*, pages 147–158, 2009.

[3] Wang De chang *et al*. Adaptive flame detection using randomness testing and robust features. *Fire safety journal 55*, pages 116–125, 2013.

[4] Töreyin B. Uğur *et al*. Computer vision based method for real-time fire and flame detection. *Pattern recognition letters*, pages 49–58, 2006.

[5] Daniel YT Chino *et al*. Bowfire: detection of fire in still images by integrating pixel color and texture analysis. *28th SIBGRAPI Conference on Graphics, Patterns and Images. IEEE*, pages 95–102, 2015.

[6] Poobalan Kumarguru and Siau-Chuin Liew. Fire detection algorithm using image processing techniques. *International Conference on Artificial Intelligence and Computer Science*, pages 160–168, 2015.

[7] Tom Toulouse *et al*. Automatic fire pixel detection using image processing: a comparative analysis of rule-based and machine learning-based methods. *Signal, Image and Video Processing*, pages 647–654, 2016.

[8] Panagiotis Barmpoutis *et al*. Fire detection from images using faster r-cnn and multidimensional texture analysis. *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8301–8305, 2019.

[9] Khan Muhammad *et al*. Efficient deep cnn-based fire detection and localization in video surveillance applications. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, pages 1419–1434, 2018.

[10] Qi xing Zhang *et al*. Wildland forest fire smoke detection based on faster r-cnn using synthetic smoke images. *Procedia engineering*, pages 441–446, 2018.

[11] Qingjie Zhang *et al*. Deep convolutional neural networks for forest fire detection. *International Forum on Management, Education and Information Technology Application. Atlantis Press*, 2016.

[12] Tom Toulouse *et al*. Computer vision for wildfire research: An evolving image dataset for processing and analysis. *Fire Safety Journal 92*, pages 188–194, 2017.

[13] Bolei Zhou *et al*. Learning deep features for discriminative localization. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016.

[14] FIREFRONT. Real-time forest fire mapping and spread forecast using unmanned aerial vehicles. www.firefront.pt/.

Real-Time 3D Tracking of Simple Objects with an RGB Camera

Student Lino Pereira[1]
lino.cp95@gmail.com
Student Bernardo Ferreira[1]
bernardoppferreira@gmail.com
Prof Alexandre Bernardino[2]
http://www.isr.tecnico.ulisboa.pt/~alex

[1] Instituto Superior Técnico, 1049-001 Lisboa, Portugal
[2] Institute for Systems and Robotics, Instituto Superior Técnico, 1049-001 Lisboa, Portugal

Abstract

This work, intends to improve a monocular region-based tracking algorithm using an RGB camera. The algorithm to be improved, derives from a particle filter where each particle represents a hypothesis of the state of the object in 3D. However, the literature mentions that the particle filter (PF) uses a very limited importance distribution to propagate the particles, which easily leads the filter to degenerate and loose track of the object. Given the limitation of the PF, an unscented particle filter (UPF) is proposed. This one obtains an approximation to the optimal importance distribution, by adding a current observation of the state.

In order to compare the proposed algorithm with the previous one, both are implemented and several real and simulated experiments with a simple object are performed. From the results obtained, is shown that the filters are successful, with the UPF being more robust.

1 Introduction

The most known methods to track an object using an RGB camera, are the methods based on 3D reconstruction and the ones based on test hypothesis. The first methods, start by using the visual information of the 2D image to reconstruct the pose of the object in 3D and the other ones, consists in generating numerous hypothesis about what could be the exact state of the object in 3D, and test each hypothesis from the 2D image information. The advantages of 3D reconstruction is that it's fast and easy to localize the object of interest, however, they are easily affected by noise in the image. On the other hand, the last methods are more precise because the image's noise is not taken into account, yet, they are very slow and poor in localizing the object. The main intent to use the UPF is to combine both methods to get the advantages of both. Thus, by defining the particles as 3D object's state hypothesis and introducing a current measurement of the object's state through a 3D reconstruction process, an hybrid algorithm is formulated.

In Bayes perspective and under the Markov assumption, the problem is to recursively estimate the posterior distribution of the current state $\mathbf{x}_t$ conditioned on all available observations $\mathbf{z}_{1:t} = \{\mathbf{z}_1, ..., \mathbf{z}_t\}$. One just needs to define some initial prior $p(\mathbf{x}_0)$, state transition $p(\mathbf{x}_t|\mathbf{x}_{t-1})$, and observation $p(\mathbf{z}_t|\mathbf{x}_t)$ probabilities, in mathematical terms [5]:

$$p(\mathbf{x}_t|\mathbf{z}_{1:t}) \propto p(\mathbf{z}_t|\mathbf{x}_t) \int p(\mathbf{x}_t|\mathbf{x}_{t-1}) p(\mathbf{x}_{t-1}|\mathbf{z}_{1:t-1}) d\mathbf{x}_{t-1}. \quad (1)$$

However, the equation (1) is intractable, and for this reason many kind of numerical approximations, like the methods based on particles, have been developed. They represent the posterior distribution as N weighted set of Monte Carlo samples $\left\{\mathbf{x}_t^{(i)}, w_t^{(i)}\right\}, i = 1, ..., N$, also known as particles, and by the law of the big numbers, the bigger the number of particles the lower is the variance of the approximation error [6]. Unfortunately, it's often impossible to sample directly from the posterior distribution, so a known and easy-to-sample distribution $q(\mathbf{x}_t^{(i)}|\mathbf{x}_{0:t-1}^{(i)}, \mathbf{z}_{1:t})$, called importance distribution is applied. By drawing samples from this distribution, a recursive estimate for the importance weights can be derived [6]:

$$w_t^{(i)} \propto \frac{p(\mathbf{z}_t|\mathbf{x}_t^{(i)}) p(\mathbf{x}_t^{(i)}|\mathbf{x}_{t-1}^{(i)})}{q(\mathbf{x}_t^{(i)}|\mathbf{x}_{0:t-1}^{(i)}, \mathbf{z}_{1:t})} w_{t-1}^{(i)}. \quad (2)$$

This type of methods exhibit a phenomenon called degeneration. This happens when some particles get all the weight and a lot of them get insignificant. To prevent this, a process of resampling is implemented to replicate the particles with high weights and discard the lower ones [6]. Doing this, brings more particles to regions of high likehood, which not only contributes to get better estimates, but also to avoid the particles from moving wrongly in the state space. One filter that derives from these type of methods, is the particle filter, that uses the simple state transition probability as the importance distribution. Yet, the literature mentions that in this type of methods, the most critical design issue is the choice of importance distribution. If the likelihood function is to narrow, or if it lies in one of the tails of the prior distribution, even the resampling process might not be enough to prevent degeneration [6]. In a Markov process, the optimal importance distribution in terms of minimizing the variance of the weights is given by:

$$q(\mathbf{x}_t|\mathbf{x}_{0:t-1}, \mathbf{z}_{1:t}) = p(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{z}_t). \quad (3)$$

However, sampling from this distribution is non-trivial, because of the dependence on the atual observation $\mathbf{z}_t$, thus, with the intent of getting an approximation to this distribution, the unscented particle filter introduced by Van Der Merwe *et al*. [6] was developed. This one uses the unscented kalman filter (UKF), which introduces a current observation together with a Gaussian approximation of the state, as the importance distribution to propagate each particle [6]:

$$q(\mathbf{x}_t^{(i)}|\mathbf{x}_{0:t-1}^{(i)}, \mathbf{z}_{1:t}) = \mathcal{N}(\mu_t^{(i)}, \mathbf{P}_t^{(i)}), i = 1, ..., N. \quad (4)$$

2 Methodology

The algorithm to be improved in this work is the one developed by M. Taiana *et al*. [5], in which to track a homogeneous ball, the state is defined as the position and velocity of the ball in space $\mathbf{x}_t = \begin{bmatrix} x & y & z & \dot{x} & \dot{y} & \dot{z} \end{bmatrix}^T$. The algorithm is based on a particle filter, where each particle represents a 3D hypothesis of the ball's state, this allows one to overcome the inversion of the nonlinearity caused by the camera projection model and enables the use of realistic 3D motion models as the state transition probability [5]. On the observation model, each particle project a few tens of points onto the current image of a video frame from his state hypothesis, one set inside and the other outside the 3D object's silhouette. With the chromatic information of these points, a normalized color histogram for the inner region and another for the outer region, are constructed along with the normalized color histogram of the object's color model [5]. The likelihood of a particle is considered high, if the inner and model histograms are similar and at the same time, the inner and outer histograms are different. To express this mathematically, a metric $\mathcal{D}$ is constructed, based on the Bhattacharyya coefficient that quantifies the similarity between histograms. At last, the observation probability of each particle is modeled by a Laplacian distribution over the metric $\mathcal{D}$, where ε was set to $\varepsilon = 1/30$ [5]:

$$p(\mathbf{z}_t|\mathbf{x}_t^{(i)}) \propto e^{-\frac{\mathcal{D}}{\varepsilon}}. \quad (5)$$

2.1 Proposed algorithm

On the previous algorithm, a motion model is applied to predict the next state of the particles, for the UPF, it's the unscented kalman filter that is used. This filter returns a prediction considering a motion model and a current observation, where the observation is a measure of the 3D position of the ball. The measurement process, consists in a method to estimate the current 3D position of the ball from an image. In order to accomplish this, a few steps must take place. The method first starts with color segmentation to identify the whereabouts of the ball. The ball is identified in the image through the highest pixel probability, corresponding to the

reference histograms created based on the ball's color model, and the image is then binarized with the use of Otsu threshold [3] to distinguish the ball from the background. Next, with the use of morphological operators, the possible noise that survived the threshold, is removed and the edges of the ball smoothed. The contour is extracted with the Moore Neighborhood [4] tracing algorithm and these points are used to extract an ellipse with the RANSAC [2] algorithm. With the best fitted ellipse, the 3D position is then obtained through monocular reconstruction [1] that uses the prior information of the ball radius and camera parameters to estimate a position from the ellipse fitted to the blob.

3 Results & Discussion

In order to access and compare the performance of the PF and UPF, several tests over simulated trajectories and real experiments were made. Results for a simulated circular trajectory and for a real free-fall trajectory are shown in this section. For both filters, there are adjustments parameters that affects the performance. The principal and only parameters tested are: the number of particles (the higher the number the better are the estimates), the distance between the inner and outer points (that controls the measurement error) and the process model noise (that regulates the dispersion of the particles in the state space). For all the plots, the tests were made using $N = 1024$ particles, where the red lines represent closer inner and outer points, the green corresponds to points a bit more distant than the red ones, and the blue even more distant. The solid, dashed and dotted lines, represent three different process noise configurations. For all tests, the motion model used corresponds to a constant velocity model. To analyze the influence of the number of particles, the root mean square error (RMSE) was used and to examine the influence of the silhouette distances and process noises, precision plots were created. This plots instead of the RMSE, can catch if a filter looses track of an object, and for filters like these, this scenario often happens. Precision plots express the percentage of estimates that possess an error below a given error threshold, as the error threshold increases. The considered error threshold is the relative error δ. Therefore, the following equation is used to compute the percentage of the estimates F, that possess an error below an arbitrary relative error δ:

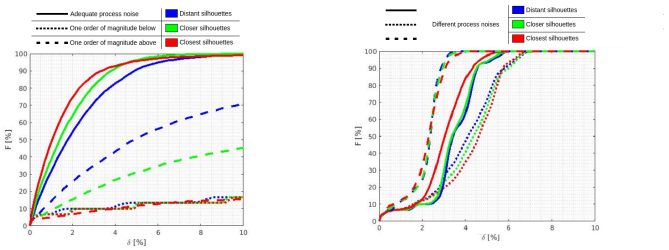
$$F = \frac{100}{N} \sum_{i=1}^{N} H\left(\delta - \frac{||\mathbf{x}_i - \hat{\mathbf{x}}_i||}{||\mathbf{x}_i||}\right) \quad [\%] \tag{6}$$

where H is the Heaviside function, N represents the number of estimates that belong to the experience, and $\mathbf{x}_i$ and $\hat{\mathbf{x}}_i$ corresponds respectively, to the exact state and the state estimate i. Both filters deal with random variables, thus, making tests with the same parameters originates different results and for this motive each test is repeated 100 times.

N	128	256	1024
PF	6.32×10^8	392.35	24.13
UPF	27.05	26.04	23.98

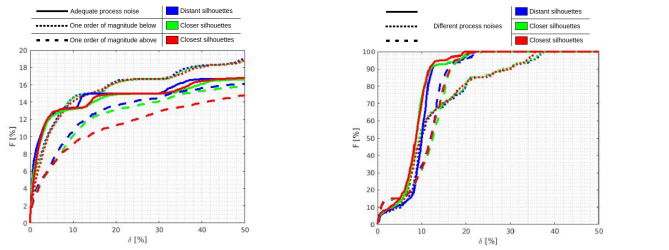
Table 1: RMSE error in *mm* using different number of particles.

In the table 1 are exposed results of RMSE of the position estimations for a given experience, varying only the number of particles. One can verify that the results coincides with the literature, once as the number of particle increases, better are the estimates, but the lower is the computational efficiency. Comparing the real results against the simulated ones (figure 1 and figure 2), it is quite visible, that for the real experiences, the relative error is bigger, at least the double, which make sense, because the simulator does not take into account the real phenomena of the world. Other aspect is the high sensitivity that the PF exhibits for different process noises and different distances between the inner and outer points (see figure 1(a)). For this filter, the process noise is directly related to the acceleration of the object and for one order of magnitude below or above the process noise used by the solid lines, the filter looses track of the ball and degenerates, which leads to wrong estimates. For the dotted lines the problem is the low scattering of the particles, and so, the filter cannot keep up with the object's movement. For the dashed lines, the particles get scattered too much and deviate from the ball which degenerates the filter. For the UPF the estimates of the position are all adequate for different process models. The real trajectory is a free-fall in which the ball collides with the ground multiple times, that makes the ball to rapidly change it's movement. That's why for the PF, the obtained results are very poor (see figure 2(a)). After an impact, the particles easily loose track of the ball and hardly recover to regions of high likelihood. On the other hand, one can see the real advantage of the UPF. If the particles loose track of the ball (mainly after an impact), the current observation acquired from a 3D reconstruction based method that easily localize the ball, will pull the particles to regions of high likelihood. Despite using such a limited motion model for this trajectory, the UPF obtains satisfactory results (see 2(b)).



(a) PF. (b) UPF.

Figure 1: Position estimates for the simulated circular trajectory.



(a) PF. (b) UPF.

Figure 2: Position estimates for the real free-fall trajectory.

4 Conclusions

The results obtained in this work, allows one to conclude that the implemented filters function with success, if the filters initial parameters are adjusted accordingly to the object's trajectory. For high uncertainty trajectories like a free-fall, the PF easily degenerates, contrarily, the UPF was successfully in all tests for any trajectory, which allows one to conclude that it's way more robust against all the three tested parameters. As future work, different observation models can be developed in order to make the algorithms usable for more complex objects.

Acknowledgements

This work was supported by:
FCT with the LARSyS - FCT Project UIDB/50009/2020.

References

[1] N. Greggio and J. Gaspar *et al*. Monocular vs binocular 3d real-time ball tracking from 2d ellipses. In *ICINCO 2011 - Proceedings of the 8th International Conference on Informatics in Control, Automation and Robotics*, volume 2, June 2011.

[2] K. Kanatani, Y. Sugaya, and Y. Kanazawa. Ellipse Fitting. In *Guide to 3D Vision Computation. Advances in Computer Vision and Pattern Recognition*, pages 11–32. Springer, Cham, 2016.

[3] N. Otsu. A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, 9(1):62–66, 1979.

[4] P. Reddy and V. Amarnadh *et al*. Evaluation of Stopping Criterion in Contour Tracing Algorithms. *International Journal of Computer Science and Information Technologies*, 3(3):3888–3894, 2012.

[5] Matteo Taiana and Joao Santos *et al*. Tracking objects with generic calibrated sensors: An algorithm based on color and 3d shape features. *Robotics and Autonomous Systems*, 58(6):784 – 795, 2010.

[6] R. Van Der Merwe and A. Doucet *et al*. The unscented particle filter. In *Proceedings of the 13th International Conference on Neural Information Processing Systems*, page 563–569, Cambridge, MA, USA, 2000. MIT Press.

Cluster-based Anchor Box Optimisation Method for Different Object Detection Architectures

Ana Filipa Sampaio
ana.sampaio@fraunhofer.pt
João Gonçalves
joao.goncalves@fraunhofer.pt
Luís Rosado
luis.rosado@fraunhofer.pt
Maria João M. Vasconcelos
maria.vasconcelos@fraunhofer.pt

Fraunhofer Portugal AICOS
Porto, Portugal

Abstract

In object detection frameworks based on deep learning, the pre-established anchor boxes are critical to ensure an adequate localisation of the objects that should be detected. As some datasets comprise objects of distinctive shapes and specific sizes, this work describes a methodology to adjust the anchor attributes to the dataset used for the task at hand.

For that, an analysis of the dataset's bounding box properties is performed, and *k*-means clustering is applied to identify the rectangular box scales and ARs that yield the best representation of the object dimensions and shapes existing in the dataset. The particularities of four popular object detection meta-architectures were taken into account to ensure that the output of the proposed method is fully compatible with the anchor box settings of different networks. The application of this methodology is illustrated using a private cervical cancer dataset.

1 Introduction

Recently, the popularity of deep learning models for object detection tasks has arisen, owing to their robustness and promising performances. These algorithms aim at localising the objects in each image in terms of rectangular bounding boxes that mark the region of the object, while also distinguishing their class [1]. Most model architectures devised for this purpose achieve the detection step through the proposal of object regions and the regression of their bounding box coordinates, with many resorting to anchor boxes, or box priors, to generate the object proposals [2, 3, 4].

Anchor boxes are bounding box templates extracted at pre-defined locations of the feature map of the convolutional neural networks (CNNs) that define the object candidates assessed by the network. Their dimensions may be directly set [5] or specified in terms of the scales and aspect ratios (ARs) combined to define the candidates extracted at each location [2, 3]. Due to their role in object proposal, anchor box settings are critical to ensure the reliable localisation of the objects in the image; hence, the anchor box scales and ARs ought to be carefully defined, bearing in mind the specificities of the annotated objects in the dataset.

Although the anchors considered in the most common object detection architectures are designed to encompass myriad object scales and shapes, in some scenarios the object proposals generated using generic anchors might not be able to match the objects that should be detected. Thus, this work presents a methodology to adjust the anchor properties to the type of objects existing in a specific dataset, enabling a more targeted object proposal procedure. Clustering is used to identify the most representative bounding box dimensions and shapes present in the dataset, which are mapped to the parameters of specific object detection CNNs, taking into account their design differences. Finally, the application of this methodology is demonstrated using a private cervical cancer dataset.

2 Methodology

The idea of exploiting dimension clusters to adjust the box priors used for object detection was already proposed in [5]. In that work, *k*-means clustering is applied to the bounding box width and height values of the training data to find several cluster centres, each associated with distinct anchor dimensions. The optimal number of anchors is established by finding the number of centres that allows a high average intersection over union (IoU) between the anchors and the ground truth boxes and does not increase substantially the computational complexity of the algorithm. The dimensions (height and width) that characterise the cluster centres are used directly to define the anchor boxes considered by YOLO.

However, in other object detection models (such as Faster R-CNN, SSD and RetinaNet), the size and shape of the anchor boxes are parameterised separately through the specification of several box scales and ARs, combined to determine the dimensions of the anchors extracted from the feature maps. Ergo, the proposed methodology applies the *k*-means algorithm in 3 distinct domains: the bi-dimensional height and width space (described above); and the domains of bounding box scales and ARs as separate variables, since this enables an easier adaptation to the way the anchor boxes are defined in the other meta-architectures. In this case, the within-cluster sum-of-squares distance metric is minimised to find the optimal clustering centres for each *k* value.

2.1 Aspect ratio and scale clustering

To find the optimal anchor shapes, the ARs of the dataset's bounding boxes are computed as the ratio between the width and the height of each bounding box. The anchor scales are computed as the ratio between the area of each bounding box and the area of the whole image. The *k*-means clustering algorithm is applied independently to the scale and AR values, finding the optimal cluster centres for each of these variables. For both properties, several *k* (number of cluster centres) values are tested and evaluated based on the sum of squared distances between each bounding box instance and its nearest cluster centre.

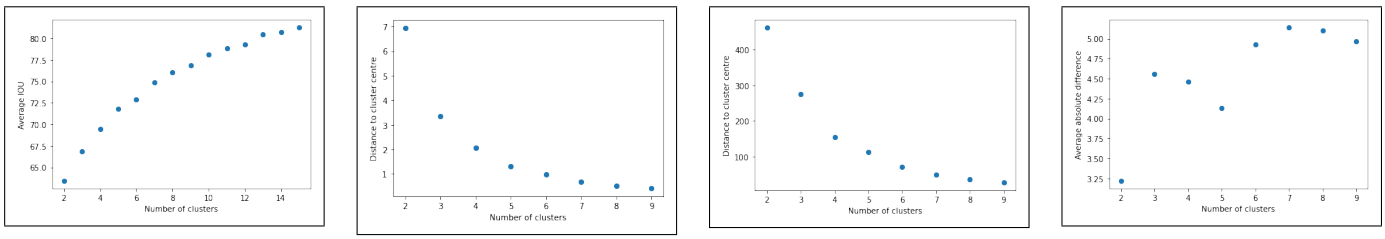
2.2 Selection of the optimal anchor scales

More cluster centres are expected to result in anchors more representative of the dimensions of the objects in the dataset, as verified in [5]; yet, the consideration of more bounding box scales and ARs implies the generation of many more object candidates during the training and execution of the algorithm, subsequently increasing its computational burden. Thus, the selection of the number of scales and ARs used in the detection algorithm is accomplished considering the **trade-off between the sum-of-squares distance** - representative of the intra-cluster variability, which should be minimised - **and the inherent computational complexity**.

In addition, the **design differences** of the current state of the art detection architectures should also be taken into account for the specification of the anchor box settings, since they might affect the anchors extracted in the object proposal step. To address these variations, four architectures - YOLO [5], Faster R-CNN [2], SSD [3] and RetinaNet [4] were examined.

One of the key disparities among these frameworks is associated with the convolutional layers from which the object proposals are retrieved: YOLO and Faster R-CNN apply the anchors to a single feature map, whereas SSD and RetinaNet propose boxes of different scales by extracting candidates from network layers of varying depths. Moreover, as aforementioned, the YOLO model contrasts with the remaining architectures by defining the anchor box dimensions directly, instead of setting the box priors through scale and AR combinations.

Even though SSD and RetinaNet both resort to feature maps of multiple depth levels to propose objects at different scales, in the SSD framework each extraction layer is associated with a single scale, whereas RetinaNet allows the specification of more than 1 sub-scale for each level. Therefore, in the SSD model, the number of feature maps used for anchor generation is equal to the number of object scales considered, and there is a direct mapping between the selected scales and the anchor extraction layers. Alternatively, in the RetinaNet framework, a feature pyramid



(a) (b) (c) (d)

Figure 1: Graphical representation of some metrics according to the k value used in the experiment: (a) average IoU between the cluster centres and the dataset's objects, for the width/height clustering; within-cluster sum-of-squares distance for the (b) scale and (c) aspect ratio values; (d) average absolute difference among the aspect ratio values in each set.

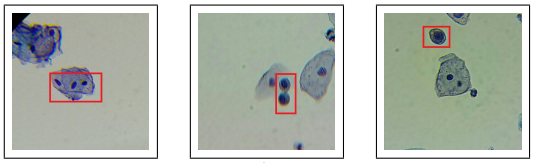
network is used to provide the convolutional layers that are the basis for anchor generation, being characterised by feature maps with a fixed consecutive resolution difference (a factor of 2), designated as octave levels. Accordingly, to take advantage of the scale values found through the proposed methodology, a careful correspondence between the selected scales and the architecture-specific parameters must be conducted.

2.3 Identification of the most discriminating aspect ratios

Given that the ARs influence the object shapes that will be more easily detected by the algorithm, the established values should be sufficiently discrepant to allow the examination of a diverse set of object shapes. To ensure this diversity, in the proposed approach, the choice of the number of ARs is based not only in the sum-of-squares distance, but also in the average absolute difference between the several AR values in each of the possible sets (inter-cluster variability).

3 Application to a private cervical cancer dataset

The approach described was applied to a private dataset comprised of 1489 microscopic images in total, acquired from liquid-based cervical cytology samples of 21 patients with a μSmartScope device [6]. This dataset includes 2436 bounding box annotations of abnormal regions (indicative of cervical lesions, illustrated in fig. 2), provided by a clinical expert from Hospital Fernando Fonseca. As the dataset had been previous split in training and test subsets according to a 80%/20% ratio, only images from the training set were analysed to establish the anchor settings.



(a) (b) (c)

Figure 2: Examples of images from the cervical cancer dataset with the bounding boxes of abnormal cells outlined (in red).

The range of k values tested in the clustering experiments was limited to a maximum of 15 for the dimensional clustering case and of 9 for the scale and AR studies to limit the number of object candidates generated. The results obtained for the several ks are depicted in fig. 1. For each k value, the final anchor dimensions, scales and ARs were obtained from the coordinates of the corresponding cluster centres.

As expected, more cluster centres lead to anchors more representative of the dimensions of the objects in the dataset, associated with a lower sum-of-squares distance error and a higher IoU metric. However, it is important to select a number of cluster centres associated with a reasonable number of anchors. Hence, adequate values for the **anchor dimensions used in YOLO** (directly defined in the image domain) would be the ones obtained for $k = 9$, for instance, resulting in the anchors of normalised dimensions $(0.36, 0.38)$, $(0.30, 0.20)$, $(0.15, 0.39)$, $(0.19, 0.19)$, $(0.3, 0.58)$, $(0.30, 0.29)$, $(0.23, 0.28)$, $(0.67, 0.67)$, $(0.57, 0.28)$. An appropriate choice for the **scale values** could be the scales of the cluster centres for $k = 6$ (0.06,0.13,0.25,0.44,0.66,0.91), since these would produce a restricted number of object proposals while keeping the same number of feature maps used in the original implementation, which is an advantage when pre-trained models are used.

The **selection of the ARs** should be grounded not only in the intra-cluster variability, but also considering how well-separated a cluster is from other clusters. Even though the ARs reported for $k = 7-9$ yielded larger differences, their consideration would increase the computational burden of the model. As the ARs clustered for $k = 6$ (0.68, 1.18, 1.90, 3.63, 7.47, 14.55) exhibit an average difference metric similar to the ones produced by more ARs, these would be suitable for the dataset analysed.

4 Discussion and conclusions

This work presents a method to optimise the localisation step in object detection networks, achieved through the adjustment of the anchor boxes' settings to the properties of the dataset used. The performed analysis addressed the factors that may influence the establishment of the anchors, in particular the similarity between the extracted anchors and the dataset's objects, the computational complexity of the model, the variety of anchor shapes and the ability to implement the anchors of choice in the existing detection models. Additionally, in studies that rely on pre-trained networks for fine-tuning, for architectures whose number of layers is directly associated with the anchor scales extracted, the number of object proposal layers should be the same as in the original model, to fully take advantage of the pre-trained weights.

Nonetheless, the experiments reported still correspond to exploratory work and further tests ought to be conducted. Future work should include the examination of the impact of the anchors' setup in the final detection performance through the comparison of the adjusted anchor settings with the default ones, as well as a characterisation of the computational burden yielded by some of the possible anchor configurations. Different clustering approaches, as well as more informative distance metrics for cluster validation, should also be explored.

Acknowledgements

This work was done under the scope of "CLARE: Computer-Aided Cervical Cancer Screening",project with reference POCI-01-0145-FEDER-028857 and financially supported by FEDER through Operational Competitiveness Program – COMPETE 2020 and by National Funds through Fundation for Science and Tecnhology FCT/MCTES.

References

[1] Z. Zhao, P. Zheng, S. Xu, and X. Wu, "Object detection with deep learning: A review,"

[2] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-CNN: Towards real-time object detection with region proposal networks,"

[3] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Fu, and A. C. Berg, "SSD: Single shot MultiBox detector," vol. 9905, pp. 21–37.

[4] T. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," pp. 1–1.

[5] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger,"

[6] L. Rosado, P. T. Silva, J. Faria, J. Oliveira, M. J. M. Vasconcelos, D. Elias, J. M. C. da Costa, and J. S. Cardoso, "μSmartScope: Towards a fully automated 3d-printed smartphone microscope with motorized stage," in *Biomedical Engineering Systems and Technologies*, Communications in Computer and Information Science, pp. 19–44, Springer International Publishing.

Assessing the Potential of Multi-view approaches in Breast Cancer Mass Detection

Eduardo Castro[1,2]
https://eduardo-castro.github.io/

José Costa Pereira[1,2,3]
jose.c.pereira@inesctec.pt

Jaime S. Cardoso[1,2]
jaime.cardoso@inesctec.pt

[1] Faculty of Engineering of the University of Porto
University of Porto
Porto, Portugal

[2] Centre for Telecommunications and Multimedia
INESC TEC
Porto, Portugal

[3] Noah's Ark Lab
Huawei
London, UK

Abstract

In the mammography exam, two views with complementary information are obtained for each breast. The state-of-the-art algorithms used in this context do not fully leverage this complementary aspect of the exam. In this work we show the potential of multiview approaches to the problem of lesion detection. For this we compare two models, one which is trained to classify between lesion and non-lesion patches and the second which, given a patch of the lesion in one view ranks candidates of that same lesion on the other view. The second model outperforms the first, showing the potential of using information from one view to guide decision for the other.

1 Introduction

Worldwide, breast cancer is the most frequently diagnosed and most lethal form of cancer in women [3]. The cumulative risk of developing the disease before the age of 75 is a little over 5%. Early diagnosis is vital in addressing this disease as it frequently translates into a better prognosis and allows more treatment options such as breast-conserving surgery [2]. Due to this, different countries implemented screening programs to anticipate detection. Screening mammography is the most commonly used imaging exam in this context and has been shown to decrease mortality [1]. Recent works have focused on developing more accurate Computer-Aided Diagnosis (CAD) algorithms [6]. The developments in deep learning in recent years and, consequently, improved image recognition models have fueled and shaped these efforts [4].

The detection of breast cancer in the screening mammography exam consists of finding lesions, often subtle, indicative of the disease. For each breast, two images are obtained for different projections (views) of the breast, which complement each other information-wise. Often, radiologists observe the same lesion in both views before making a decision.

The state-of-the-art algorithms for breast cancer screening do not integrate the information at the lesion level. They fuse the knowledge of the two views either by averaging the "diagnosis" or aggregating image-level features. Lesion level integration has the potential to improve accuracy and the interpretability of the results returned by the algorithm. This work is a starting point in this direction. We show that a lesion's information in one view is useful to detect the lesion on the other view.

2 Methods

2.1 Baseline: Image Classification with CNNs

Convolutional Neural Networks (CNNs) are the most common type of neural network in vision applications. In recent years researchers have adopted these models in the context of Computed Aided Diagnosis tools for Breast Cancer screening. Image classification is commonly done by minimization of the cross-entropy loss function on a training set:

$$\mathcal{L}_H = \sum_{c=1}^{M} y_c \log(p_c) \tag{1}$$

where $y_c \in \{0,1\}$ is 1 if the label of the image is class c, and p_c is the probability assigned by the model that the image belongs to class c.

2.2 Multiview: An approach based on the Triplet Loss

In this work, we considered an alternative setting for image classification in which the patch corresponding to the lesion in the other view (anchor) is given. In this context, the model can obtain information on what the lesion might look like. The triplet loss [7] can thus be used to train a model that tells us if there is a correspondence between the anchor and the candidate. This loss function is given by:

$$\mathcal{L}_T = \max(d(a,p) - d(a,n) + margin, 0) \tag{2}$$

where a, p, n are feature representations for the *anchor*, the *positive* and the *negative* images and d is a measure of distance (euclidean norm in the case of this work). The minimization of this loss function leads to the desirable case where: $d(a,p) < d(a,n) + margin$. Thus, from all candidates it is expected that the patch that minimizes $d(a,x)$ is the correct one.

2.3 Hybrid: Aggregating the two decisions

The two proposed models are conceptually different. While the baseline learns to discriminate between positive and negative patches, the multiview approach relies on the anchor. As such, the information they base their decisions on may be complementary.

A third option is to use a hybrid strategy that relies on the decision of the two models. Here we propose a simple rule in which the final score, which ranks the candidates, is obtained with the **baseline** model plus a Δ if that candidate is the preferred one for the **multiview** model.

3 Experiments and Discussion

Due to its size and accessibility, CBIS-DDSM [5] is the leading publicly available dataset for developing breast cancer screening algorithms. This collection is an updated and standardized version of DDSM and contains approximately 10k images. Each finding in the dataset is associated with a segmentation mask and its pathology (malign or benign). Images were obtained from scanned film mammography. In this work, a subset of this data with around 1200 images was used.

The data was split at the patient level into three sets: train (70%), validation (10%), and test (20%). Positive patches were taken centered on each lesion's mask. A custom lesion detector was employed to obtain five false negatives per image in the dataset to serve as additional candidates. All patches were resized to 64×64.

A custom architecture was used for all experiments with eight convolutional and two fully-connected layers. The models were trained for around 80k iterations with a starting learning rate of 0.01, which was decreased one time by a factor of 10, using stochastic gradient descend with momentum, with a batch size of 32. Batch normalization and weight decay were used. Each experiment was repeated five times.

Each method's accuracy was computed by first ranking the candidates and then selecting their top choice for each image. This selection is considered correct if it corresponds to the lesion and incorrect otherwise. The top candidate for the **baseline** model was the one that maximized the probability of being positive while for the **multiview** model, the one that minimized the distance to the *anchor*. The Δ for the hybrid strategy was 0.25. Results are shown on table 1. In Figure 2, the sensitivity per average number of false positives is shown for the three models.

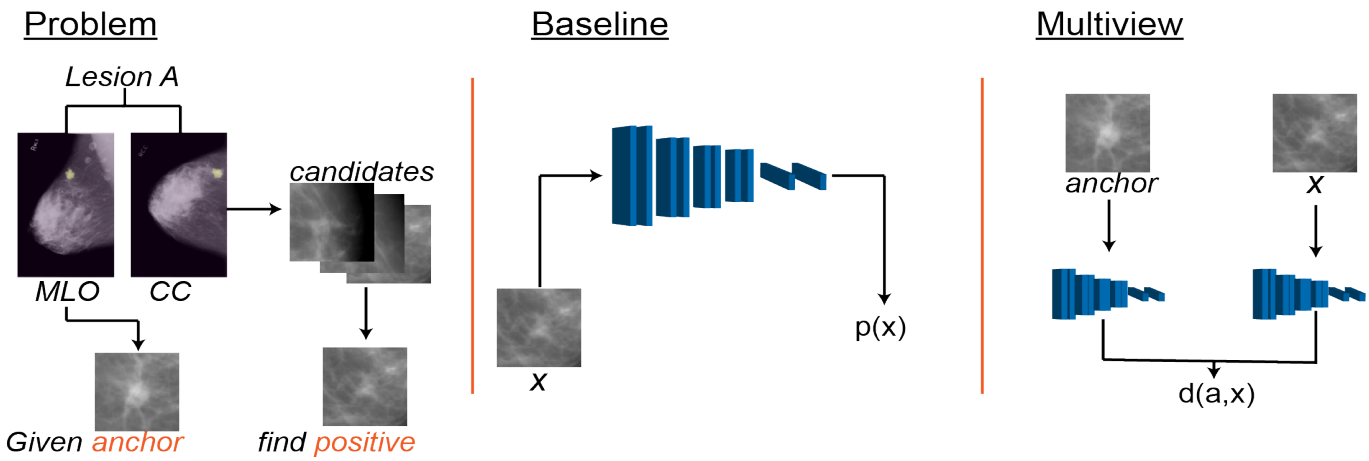


Figure 1: Diagram for the problem formulation and strategies followed in this work.

Table 1: Test set accuracy for each method.

Method	Baseline	Multiview	Hybrid
Accuracy (%)	76.13 ± 1.64	80.4 ± 0.6	82.02 ± 1.62

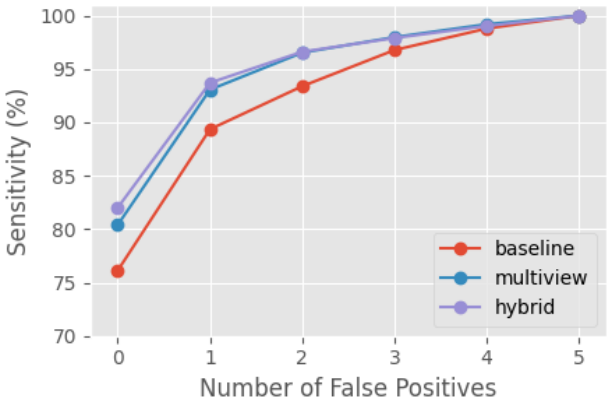


Figure 2: Sensitivity per false positive.

As shown, the **multiview** approach outperforms the baseline, showing that the appearance of the anchor image can be used to enhance the detection of the same lesion on the other view. When combined in the hybrid strategy, the two models can enhance one another, suggesting that their information is complementary.

Even though the results show the importance of using a **multiview** approach at the lesion detection level, future research is needed on how to integrate the two losses together in a single training scheme. Also, it is desirable to have an automatic algorithm that does not require an *anchor*. In this context, it is yet to be shown the value of a **multiview** approach when there are only candidates for view, rather than a "known" positive.

4 Conclusions

Breast cancer is a considerable burden on patients worldwide. The development of more accurate CAD systems can help reduce this burden through earlier and more accurate diagnosis. The development of CNNs has allowed an increase in accuracy. However, current methods could be further improved by better integrating the information of the two views. This work demonstrates this potential by showing that a model that uses the opposite view's appearance can outperform a naive baseline in lesion detection. Future research should focus on how to translate this potential to a more realistic/less controlled scenario.

5 Acknowledgements

The project TAMI - Transparent Artificial Medical Intelligence (NORTE-01-0247-FEDER-045905) leading to this work is co-financed by ERDF - European Regional Fund through the Operational Program for Competitiveness and Internationalisation - COMPETE 2020, the North Portugal Regional Operational Program - NORTE 2020 and by the Portuguese Foundation for Science and Technology - FCT under the CMU - Portugal International Partnership. This work is also financed by National Funds through the Portuguese funding agency, FCT - Fundação para a Ciência e a Tecnologia, within PhD grant number *SFRH/BD/136274/2018*. The authors would also like to aknowledge NVIDIA for their generous donation of a TitanX gpu.

References

[1] Donald A. Berry, Kathleen A. Cronin, Sylvia K. Plevritis, Dennis G. Fryback, Lauren Clarke, Marvin Zelen, Jeanne S. Mandelblatt, Andrei Y. Yakovlev, J. Dik F. Habbema, and Eric J. Feuer. Effect of screening and adjuvant therapy on mortality from breast cancer. *New England Journal of Medicine*, 353(17):1784–1792, 2005. doi: 10.1056/NEJMoa050518. PMID: 16251534.

[2] Carol E. DeSantis, Jiemin Ma, Mia M. Gaudet, Lisa A. Newman, Kimberly D. Miller, Ann Goding Sauer, Ahmedin Jemal, and Rebecca L. Siegel. Breast cancer statistics, 2019. *CA: A Cancer Journal for Clinicians*, 69(6):438–451, 2019. doi: 10.3322/caac.21583.

[3] J. Ferlay, M. Colombet, I. Soerjomataram, C. Mathers, D.M. Parkin, M. Piñeros, A. Znaor, and F. Bray. Estimating the global cancer incidence and mortality in 2018: Globocan sources and methods. *International Journal of Cancer*, 144(8):1941–1953, 2019. doi: 10.1002/ijc.31937.

[4] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, May 2015. ISSN 0028-0836. doi: 10.1038/nature14539.

[5] Rebecca Lee, Francisco Gimenez, Assaf Hoogi, Kanae Miyake, Mia Gorovoy, and Daniel Rubin. A curated mammography data set for use in computer-aided detection and diagnosis research. *Scientific Data*, 4:170177, 12 2017. doi: 10.1038/sdata.2017.177.

[6] Thomas Schaffter, Diana S. M. Buist, Christoph I. Lee, Yaroslav Nikulin, Dezső Ribli, Yuanfang Guan, William Lotter, Zequn Jie, Hao Du, Sijia Wang, Jiashi Feng, Mengling Feng, Hyo-Eun Kim, Francisco Albiol, Alberto Albiol, Stephen Morrell, Zbigniew Wojna, Mehmet Eren Ahsen, Umar Asif, Antonio Jimeno Yepes, Shivanthan Yohanandan, Simona Rabinovici-Cohen, Darvin Yi, Bruce Hoff, Thomas Yu, Elias Chaibub Neto, Daniel L. Rubin, Peter Lindholm, Laurie R. Margolies, Russell Bailey McBride, Joseph H. Rothstein, Weiva Sieh, Rami Ben-Ari, Stefan Harrer, Andrew Trister, Stephen Friend, Thea Norman, Berkman Sahiner, Fredrik Strand, Justin Guinney, Gustavo Stolovitzky, , and the DM DREAM Consortium. Evaluation of Combined Artificial Intelligence and Radiologist Assessment to Interpret Screening Mammograms. *JAMA Network Open*, 3(3):e200265–e200265, 03 2020. ISSN 2574-3805. doi: 10.1001/jamanetworkopen.2020.0265.

[7] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. *CoRR*, abs/1503.03832, 2015.

Object Detection in Equirectangular Images

Francisco Henriques[1]
2180302@my.ipleiria.pt

Joana Costa[1,2]
joana.costa@ipleiria.pt

Catarina Silva[2]
catarina@dei.uc.pt

Pedro Assunção[1,3]
pedro.assuncao@ipleiria.pt

[1]ESTG, Polytechnic of Leiria, Leiria, Portugal

[2] Universidade de Coimbra, CISUC - Centro de Informática e Sistemas, FCTUC-DEI - Departamento de Engenharia Informática, Portugal

[3]Instituto de Telecomunicações, Leiria, Portugal

Abstract

Nowadays, computer vision (CV) is widely used to solve real-world problems, which pose increasingly higher challenges. In this context, the use of omnidirectional video in a growing number of applications, along with fast development of Deep Learning (DL) algorithms for object detection, drives the need for further research to improve existing methods specifically developed for conventional 2D planar video. This work explores DL methods to detect visual objects in omnidirectional images represented onto plane through Equirectangular Projection (ERP). It is shown that the error rate of object detection using existing DL models with ERP images depends on the object spherical location in the image. Then, a new object detection framework is proposed to obtain uniform error rate across the whole spherical image regions.

1 Introduction

Over the last decades, computer vision (CV) technology, through traditional or intelligent approaches, has been widely explored to solve real-world problems through advanced technology in many different domains, such as self-driving cars, accurate health diagnoses, agriculture operations improvement, remote surveillance and monitoring, etc [1]. Such systems are usually based on planar images captured from 2-dimensional (2D) cameras, usually referred to as conventional cameras. However, new application requirements and fast technological advances are continuously posing new challenges which cannot be met by conventional cameras. Their limited field-of-view (FOV) and, subsequently, blind spots do not allow all view directions, including all-around from the ground, mid-level above ground to sky, to be monitored. For instance, in outdoor smart surveillance systems, a conventional camera no longer meets the requirements posed by all types of possible intrusions in private properties or high security areas. In fact, nowadays intrusion may happen either physically at the front door or remotely through a flying drone. To cope with such new demands, omnidirectional vision has been evolving in several directions such as: object detection and identification, people recognition, vehicles traffic monitoring, etc., at the ground-level; monitoring buildings, balconies, or windows at the mid-level; detect sky-level objects such as unmanned aerial vehicles (UAVs), which consist of autonomously or remote-controlled vehicles to fly over target areas .

Deep Learning (DL) approaches have been heavily studied in the last few years and nowadays there are several frameworks capable of providing reasonable performance in many image and video processing tasks. However, currently available DL frameworks were designed to use 2D data as input, while specific solutions for omnidirectional video are still open for further improvement and performance optimisation. This paper is motivated by this technological context, addressing performance optimisation of DL approaches for object detection in omnidirectional images representing the spherical domain as planar images through the well-known Equirectangular projection (ERP). The equirectangular projection defines each sphere point by a horizontal angle $\theta \in [-\pi, \pi[$ and vertical angle $\theta \in [-\pi/2, \pi/2[$. Then, given a sphere $\sum$, an Equirectangular image P is obtained by sampling the spherical surface as follows [2]:

$$P(i,j) = \sum(\theta_i, \phi_j)$$

$$\text{with } \forall_i,\ \theta_i - \theta_{i+1} = \delta\theta \text{ and } \forall_j, \phi_j - \phi_{j+1} = \delta\phi$$

Although ERP has become a popular representation format to store and transmit omnidirectional or 360º video content, it produces significant geometric distortions in regions near the poles due to non-uniform sampling density, which results from equal distances in the visual scene being represented by a different number of equally spaced pixels. Thus, the aspect ratio of the any object depends on its spherical position which makes object detection harder to achieve. Regarding the use of DL based approaches, in addition to the above-mentioned challenges, the lack of ERP labeled image datasets leads to an effort to be made by researchers to construct a decent dataset in terms of size, annotation richness, and scene variability and complexity [3].

In this paper, we show how to overcome the above-mentioned problems in a DL-based object detection framework using Equirectangular images. Firstly, a dataset acquisition stage along with the description of the steps required to reach the final dataset is described for better understanding the input of the proposed framework. Afterwards, we benchmark algorithms' performance on conventional and ERP datasets to identify the main problems concerning those techniques. Finally, a framework which allows object detection tasks to provide improved results is described in detail.

2 360º Image Dataset

In the dataset acquisition process, the first step consisted of contributing to decrease the lack of labelled Equirectangular images. For that purpose, a 360º video camera was used to capture an urban environment to include different visual objects of all possible regions of spherical images in the dataset. To that purpose, the camera was firstly placed on a highly congested traffic locations to produce video recordings where people and vehicles were visible. Then, to enrich the dataset with high diversity viewpoints, object poses, and weather conditions, the same camera was mounted on the roof of a car, and videos were recorded while the car was moving. Finally, to fill the lack of aerial objects an unmanned aerial vehicle was controlled over pre-defined regions, simulating aerial intrusion in a private property, while the 360º camera was recording playing the role of an omnidirectional surveillance camera. Afterwards, the resulting video shots were processed to extract the most representative ERP video frames, originating a total of 779 omnidirectional ERP images that were labelled using an annotation tool to identify object classes and locations considering the following class labels: car, truck, bus, motorcycle, person, and unmanned aerial vehicle.

2.1 Reference Performance on 360º Dataset

After the 360º dataset acquisition stage, a reference performance evaluation of currently available deep learning (DL) networks was carried out, using conventional planar images with small FOV when compared with 360º images. Since the proposed test experiment required a conventional image labeled dataset, we investigated open-source available datasets related to urban environment. Among the wide range of available datasets that were found, the Cityscapes [4] dataset was chosen, due to its huge diversity and application scenarios.

Therefore, taking as input, part of the Cityscapes dataset and preserving its primary organization (training, validation, and testing subsets), DL algorithms were trained through transfer-learning techniques to compare their performance on Cityscapes dataset and on the ERP dataset acquired in the scope of this work.

Reference performance experiments consisted of training Single-Shot Detection (SSD) [5] and You Only Look Once (YOLO) v3 [6] networks on the conventional image dataset (Cityscapes) and compare the resulting accuracy performance on both datasets. Considering that Cityscapes subset does not contain all object classes covered by our 360º image dataset, we have only included results for car, bus, and person labels.

		AP@0.5 (%)			mAP@0.5 (%)
		car	bus	person	
Cityscapes subset	SSD	73.2	65.8	74.3	71.1
	YOLOv3	76.3	67.1	75.3	72.9
360º dataset	SSD	47.1	28.3	41.5	39.0
	YOLOv3	49.6	30.1	44.7	47.7

Table 1: Accuracy of DL algorithms trained on conventional image dataset, measured on conventional and 360º dataset.

Despite the fact that accuracy significantly decreases from conventional to 360º dataset (as expected), both algorithms have demonstrated more difficulty to detect objects near the image centre than elsewhere (e.g. accuracy differences up to 40% were found between centre regions and others. Figure 1 depicts an example of a car located in the mid-region, which was not detected as the remaining objects.

Figure 1: Predictions in ERP images. DL algorithms show more difficulty detecting objects in centre regions.

Hence, we have considered the whole 360º dataset to evaluate the False Positives (FP) rate by image region. We have noticed that the described metric does not follow a uniform pattern, with higher values (63%) in the centre of the image than both left and right regions (16% and 21%, respectively). Given the reference performance above, the main drawbacks are identified as the lack of accuracy of existing models and the correlation between non-detected objects and image regions.

2.2 360º Dataset Training

To tackle the previous limitations domain-specific training with data augmentation approaches was carried out. We used a set of DL algorithms applied to our 360º dataset to detect cars, UAVs, and people, including two variations of YOLOv4, YOLOv3 and tiny-YOLO on both versions (3 and 4). Moreover, two variations of SSD and Mask R-CNN [7] were also evaluated.

The benchmarking analysis focuses on providing a detailed evaluation of the trained models, taking into consideration three fundamental performance metrics: mean average precision (mAP), to evaluate models' accuracy, floating-point operations per second (FLOPs), considering the computational cost associated with each deep neural network, and, finally, the model complexity, given by the number of learning parameters. Each model inference speed has also been computed by measuring the elapsed time between receiving an image and when predictions are available.

DL Network	*Parameters*	*G-FLOPs*	*mAP*	*Inference Time (ms)*
Mask R-CNN	250	**628,94**	**89**	**2011 ± 4,23**
Standard YOLOv4	244	127,294	86	349 ± 5,83
YOLOv4 - 800x448	244	123,416	82	403 ± 5,95
Standard YOLOv3	235	139,558	80	398 ± 6,41
SSD 512x512	**286**	163,262	73	451 ± 8,23
Tiny-YOLOv4	22	6,793	65	**171** ± 3,21
SSD 300x300	97	56,452	61	220 ± 5,46
Tiny-YOLOv3	**33**	**5,454**	**59**	193 ± 2,98

Table 2: Results of DL algorithms trained on 360º images.

Results presented in Table 2 demonstrate great improvements in terms of accuracy on detecting objects in ERP images compared to the same algorithms trained on Cityscapes subset (Table 1). However, the same experiment to provide FP rate by image region applied to these models has shown that this framework does not allow to meet high-accuracy requirements of most demanding applications.

3 Proposed Approach

The proposed framework consists of adding a pre and post-processing stage to the default object detection framework, which provides predictions just taking an image as input. Due to the fact that objects located at the center tend to be smaller, which could crucial to justify different FP rates, we include two pipelines: one focusing the whole image, and another just concentrating on the middle region. To perform the second pipeline, we divide the middle region into two sub-regions, as depicted in Figure 2.

To evaluate framework's performance standard YOLOv4 was used as DL algorithm on both pipelines. Then, the resulting predictions from 360º dataset inference, were, successively, compared to the labelled objects to produce the final results. Although the measured inference time has increased, mid-level FP rate have demonstrated improvements, which leads to a more uniform FP rate by image region. Measured values have shown the referred metric has decreased from 63%, in the initial experiments, to 39% on the proposed framework.

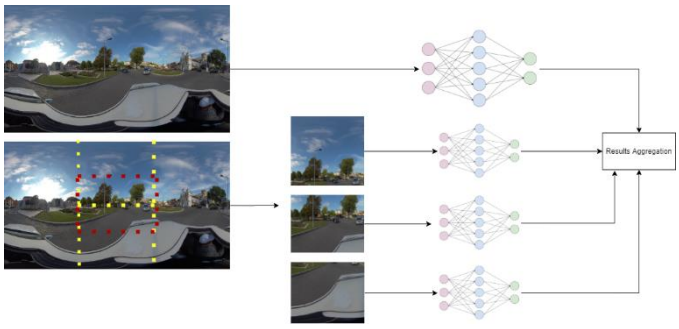


Figure 2: Proposed framework architecture with two stages. Results are aggregated with predictions from mid-region sub-divisions.

4 Conclusion

Automatic object detection in ERP images with high-level accuracy created new problems that did not occur before in conventional images. Object distortion and unusual view pose as well as very-high image resolution tend to give rise to an extremely wide range of objects dimensions and aspect ratios across an image. Our initial experiments have demonstrated that a conventional framework does not provide uniform accuracy results across the whole image. The framework proposed in this paper allows to make non-detected objects by image region more uniform through two parallel pipelines: one for the whole image and the other focusing on the most problematic region, the center.

Acknowledgments: This work was partially supported by project ARoundVision CENTRO-01-0145-FEDER-030652.

References

[1] Q. Wu, Y. Liu, Q. Li, S. Jin e F. Li, "The application of deep learning in computer vision," *Chinese Automation Congress (CAC)*, nº 17469740, pp. 6522-6527, 2017.

[2] Maugey¸Thomas, O. L. Meur e L. Zhi, "Saliency-based navigation in omnidirectional image," *IEEE 19th International Workshop on Multimedia Signal Processing (MMSP)*, nº 17411746, pp. 1-6, 2017.

[3] W. Yang, Y. Qian, J. Kämäräinen, F. Cricri e L. Fan, "Object Detection in Equirectangular Panorama," *2018 24th International Conference on Pattern Recognition (ICPR), Beijing*, nº 18303181, pp. 2190-2195, 2018.

[4] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth e B. Schiele, "The Cityscapes Dataset for Semantic Urban Scene Understanding," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[5] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu e A. C. Berg, "SSD: Single Shot MultiBox Detector," *Lecture Notes in Computer Science*, pp. 21-37, 2016.

[6] Redmon J, S. Divvala, R. Girshick e A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 779-788, 2016.

[7] H. Kaiming, G. Georgia, D. Piotr e G. Ross, "Mask R-CNN," *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 2980-2988, 2017.

Corpus Callosum Segmentation using UNET and Transfer Learning

Afonso Pinto[1]
afonsoxp@ua.pt
Regina Oliveira[1]
regina.oliveira@ua.pt
Ana Tomé[2]
ana@ua.pt
Augusto Silva[2]
augusto.silva@ua.pt

[1] Physics Department
University of Aveiro
Aveiro, Portugal

[2] Institute of Electronics and Informatics Engineering of Aveiro
University of Aveiro,
Aveiro, Portugal

Abstract

The Corpus Callosum is an important brain structure, whose function is to interconnect the brain's hemispheres. The segmentation of this structure is very challenging, but nowadays several automatic strategies to achieve this goal already exist. In this paper it will be presented a deep learning algorithm for the Corpus Callosum segmentation, using a U-Net model. Also, in this work, a transfer learning approach was performed, where the network was trained to execute the cerebellum segmentation, and the net weights were stored to apply them to the Corpus Callosum segmentation task.

The obtained results were very satisfying, achieving an average dice score of 62.51% and 81.62% for the control and the autistic patients group, respectively, making this methodology very interesting for Corpus Callosum segmentation in diagnosis tasks, for example.

1 Introduction

The Corpus Callosum (CC) is a brain structure composed of white matter, which connects the left and right brain hemispheres, being responsible for the communication between them. This structure can reach approximately 10 cm of length and 1 cm of width, containing about 200 million axonal projections [1]. Structural features of CC, like size and shape, are correlated to neurological diseases, such as epilepsy, autism, schizophrenia, and dyslexia, for example. Thus, automatic and precise segmentation can be advantageous for the diagnosis of these diseases, based on quantitative CC features [2].

When compared to manual segmentation, an automatic approach is easier to perform, saving time, and the segmentation result is independent of errors inherent to human performance. Manual segmentation of the CC is difficult by the fact that the fornix and the nervous tissue's intensity around the CC on MRI (Magnetic Resonance Imaging) images is very similar to the CC's intensity [2].

The U-Net is a convolutional neural network used recently for biomedical images segmentation purposes. This specific network is composed by a contracting (down sampling) and an expanding (up sampling) path, symmetrically placed. The first one has the architecture of a common convolutional network, composed by repeated perform of two 3x3 convolutions, each pursued by a ReLU (Rectified Linear Unit) and a 2x2 max pooling operation, responsible for the input images down sampling. The function pooling has 2 as stride. In the down sampling path, the quantity of feature channels is doubled at each step. On the other hand, the expansive path is responsible for the up sampling of the feature map at each stage followed by 2x2 convolution that reduces the feature channels to half. The convolution output is concatenated with the correspondent feature map on the same level in the descent path, reincluding the localization information, and two 3x3 convolutions are processed, each one followed by a ReLU. In the final layer a 1x1 convolution is applied to map the resulting feature vector, formed by 64 components [3].

The U-Net is a good option for segmentation goals because it is capable of combining localized and contextual information, given by the down sampling and the up sampling paths, respectively, making this network more precise when compared with others, and doesn't need a large amount of data for the training task, making use of data augmentation. Also, the use of a weighted loss function allows an accurate diagnosis separating efficiently two objects of interest, since in the training task the network gives more weight to the pixels between the objects as the distance between them decreases [3].

In this work, transfer learning was used to facilitate the learning process for the CC's segmentation, obtaining a more accurate and faster segmentation. Transfer learning approaches allow the use of less labelled training data. Succinctly, a network is previously trained for a different segmentation task, in this paper the U-Net was trained with cerebellum images, well segmented by the VolBrain platform [4], and then the knowledge (features, weights) acquired are used on the contracting path, only being necessary the training of the expanding path for the final goal, the CC's segmentation in this specific case. Some of the obtained images using this approach can be seen in Figure 1.

2 Methodology

The U-Net described in the introduction was trained using images from ABIDE dataset [5]. More specifically 32 images from the California Institute of Technology were used for training, of which 19 were of autistic patients and the rest were from control cases and 6 were used for validation of the model. The datasets partitions for training and validation were chosen randomly. For testing the trained network, 5 images of normal and 5 of autistic patients were randomly used, from the Carnegie Mellon University image repository.

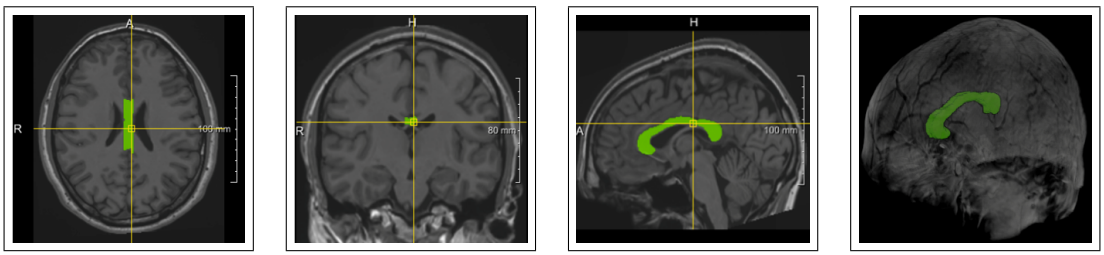
The images were all passed through the VolBrain MRI image pipeline [4] to obtain MRI scans all on the same position and to obtain cerebellum segmentations. To achieve CC segmentation, that will be used for model optimization, the software ITK-SNAP [6] was used to segment each image manually, since VolBrain does not segment this structure. Each slice of the processed volumes was padded with zeros to reach a dimension of 256x256 pixels in width and length. This last procedure was necessary to conform to the dimension requirements of the input image to the U-Net.

The training of the network was carried out in two stages. In the first one the model was fed MRI images slices as an input and cerebellum segmentations as an output, as a way to train the encoder of the model. After completion of the first stage, the weights of the encoder's layers were locked to speed up training of the following part. On the second stage, the model was trained using the same MRI images, as an input, and CC segmentations as the desired output. Cerebellum segmentations were chosen for this transfer learning method, because they represent a task that is similar to the CC segmentation (same domain).

Each stage was trained using an Adam optimizer with a learning rate of 0.0001 and a binary cross-entropy function. The training was constituted by 10 epochs, which were carried out in each stage of training. The data fed to the network was augmented through random rotation operations (in a range of 10 degrees), random horizontal and vertical flips and random shifts (in a range 0.1*image size) in the original image.

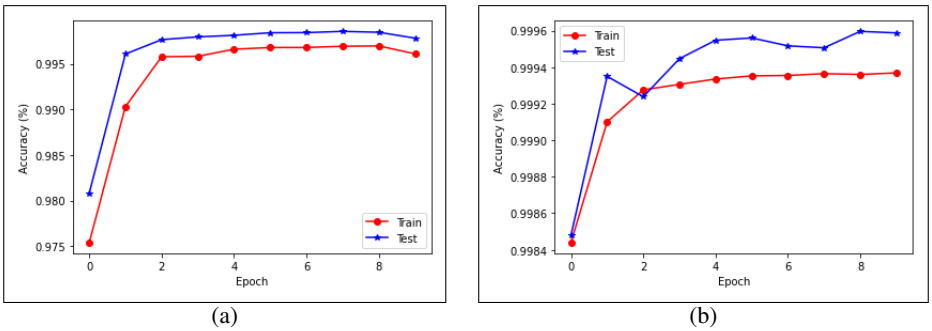
3 Results and Discussion

The training of this type of networks usually is performed using GPU. However, it was ran in CPU, due to hardware limitations, taking more time to execute the training task. For the cerebellum segmentation, the 10 epochs were performed in 21 hours and 53 minutes. For the CC segmentation, the 10 epochs were performed in 22 hours and 35 minutes. The accuracy per epoch obtained in training of both phases can be observed



(a) (b) (c) (d)

Figure 1: CC segmentation in the anatomical planes: (a) axial; (b) coronal; (c) sagittal; (d) 3D representation;



(a) (b)

Figure 2: Voxelwise accuracy evolution in the training task through epochs: (a) cerebellum segmentation; (b) CC segmentation;

on Figure 2. After training, the automatic segmentation of each volume could be completed within 45 seconds average.

Anyone seeking to reproduce the obtained results in this article should try to use a GPU for training, since the process should become much more time efficient, making it possible to run more epochs for a better understanding of the model. However, from the training history it is possible to understand that both steps of training stabilize the segmentation accuracy after 4 epochs for the training and testing datasets. This is not verified on the testing dataset of the CC segmentation, where the accuracy fluctuates a bit, approximately 0.0002%, so it was considered irrelevant. Such a phenomenon may be explained by the random errors performed in manual segmentation, that make it harder for the model to properly adapt, which can also explain why the second stage of segmentation took longer than the first one. With the use of transfer learning to segment the CC it is possible to achieve higher accuracy values in less epochs, as can be seen in Figure 2 (b), due to the fact that most of the training was already done with the cerebellum segmentation dataset.

To evaluate the network performance, 5 MRIs of normal patients and 5 MRIs of autistic patients were automatically segmented, being one example of this procedure represented in Figure 1, and compared against the manual ones. To analyze the match of both segmentation a dice score evaluation metric was used, due to its relevance in segmentation tasks. The mean results for this metric was 62.51% for the control group and 81.62% for the autistic patients group.

The achieved results for dice score were satisfactory for both groups, however the algorithm was more successful in the segmentation of brains of autistic patients. The differences may be explained by different factors. The first one is related to the fact that the CC segmentation were done manually by an untrained researcher, and as a result they are bound to have many small random mistakes in them. In fact, in some separate cases, the automatic segmentation seems better than the manual one, meaning that the poor dice score achieved is not a good measurement of the quality of segmentation. The second one is related to poor MRI image quality after VolBrain treatment. Specifically one of the images used in the control group was distorted, which caused the algorithm to falsely label a region as CC, causing the drop in the mean dice score. Perhaps more vast image augmentation procedures could be applied, to ensure that the process remains robust, even when faced with this type of problems. On other note, the training accuracy was significantly higher than in these test volumes because the metric used was different due to the limitations of the algorithm used.

4 Conclusion and Final Remarks

This algorithm was able to perform a good CC segmentation, obtaining satisfactory results in the dice score evaluation metric. However, some improvements can be done in the training task, like the execution of a manual segmentation of CC by a specialist of the area, and the application of more image augmentation procedures to increase the robustness of the algorithm, leading to better dice score results.

The use of transfer learning had various advantages along this work. First, this methodology was able to perform a good CC segmentation even with the presence of some errors in the ground truth, since it was executed by an untrained individual. Second, by using this methodology it is possible to achieve accurate segmentations of the desired anatomical structure within few epochs. If a bigger dataset is used in the initial phase, it should be possible to achieve even better results, easing the training for the end user, allowing him to obtain a segmentation tool with a smaller dataset and in shorter time.

The obtention of CC segmentation by an automatic system, like the one that was described in this paper, can be used to support diagnosis tasks, like the diagnostic of Autism Spectrum Disorder. Also, this algorithm can be used to obtain an automatic segmentation of diverse structures, giving the possibility to obtain batches of data, which can be used to study relevant properties (texture, edges, volume, etc.) of those structures, allowing the progress of several science fields.

References

[1] "Corpus Callosum". In: (2014). Ed. by Michael J. Aminoff and Robert B. Daroff, pp. 867–868. DOI: https://doi.org/10.1016/B978-0-12-385157-4.01137-4.

[2] Gilsoon Park et al. "Automatic segmentation of corpus callosum in midsagittal based on Bayesian inference consisting of sparse representation error and multi-atlas voting". In: *Frontiers in neuroscience* 12 (2018), p. 629.

[3] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. "U-net: Convolutional networks for biomedical image segmentation". In: *International Conference on Medical image computing and computer-assisted intervention*. Springer. 2015, pp. 234–241.

[4] José V. Manjón and Pierrick Coupé. "volBrain: An Online MRI Brain Volumetry System". In: *Frontiers in Neuroinformatics* 10 (2016), p. 30. ISSN: 1662-5196. DOI: 10.3389/fninf.2016.00030. URL: https://www.frontiersin.org/article/10.3389/fninf.2016.00030.

[5] Adriana Di Martino. *ABIDE - Autism Brain Imaging Data Exchange*. 2017. URL: http://fcon_1000.projects.nitrc.org/indi/abide/ (visited on).

[6] Paul A. Yushkevich et al. "User-Guided 3D Active Contour Segmentation of Anatomical Structures: Significantly Improved Efficiency and Reliability". In: *Neuroimage* 31.3 (2006), pp. 1116–1128.

Extremely Randomised Trees for Computational Complexity Reduction of Omnidirectional Intra Video Coding

Jose N. Filipe[1]
jose.filipe@av.it.pt
J. Carreria[1,2]
jcarreira@co.it.pt
Luis M. N. Tavora[2]
luis.tavora@ipleiria.pt
Sergio M. M. Faria[1,2]
sergio.faria@co.it.pt
Antonio Navarro[1,3]
navarro@av.it.pt
Pedro A. A. Assuncao[1,2]
amado@co.it.pt

[1] Instituto de Telecomunicações, Campus Universitário de Santiago Aveiro, PT
[2] Politecnico de Leiria, ESTG, Morro do Lena - Alto do Vieiro, Leiria, PT
[3] Universidade de Aveiro, Campus Universitário de Santiago, Aveiro, PT

Abstract

This paper presents a novel method to reduce the computational complexity of intra-coded 360° video in Equirectangular Projection (ERP) format. The proposed method is based on three Extremely Randomised Trees models to predict the maximum partition depth that should be used for intra-coding of the complex nested data structures (Quad Tree, Binary Tree and Ternary Tree) used in the forthcoming video coding standard (Versatile Video Coding). The results show that an average complexity reduction of 31.35% is achieved, with a negligible loss of coding efficiency of just 0.42%, outperforming other state of the art low complexity solutions.

1 Introduction

Image and video data represent the majority of all internet traffic, due to new applications and emerging services using mixed reality, namely 4K and 8K resolutions, cloud gaming, self-driving vehicles, smart surveillance systems, among others. These advanced services and applications require very high resolutions and complex compression algorithms. Thus, standardisation efforts specifically targeting emerging video formats, such as 360° video [5] are in place. Specifically, the Joint Video Exploration Team (JVET) is developing the forthcoming video compression standard, named Versatile Video Coding (VVC), to face the challenging requirements posed by higher resolutions and new visual representation formats. This new standard greatly increases the coding efficiency of its predecessor High Efficiency Video Coding (HEVC). However, such improvement is achieved at the cost of a great deal of additional computational complexity. As the VVC encoder is 7 to 9 times more complex than HEVC [10], fast computational methods are of utmost importance to ease adoption of the this standard and to meet implementation constraints.

Historically, previously proposed methods to reduce the complexity of video encoders focus on reducing the number of tests performed by the Rate-Distortion Optimisation (RDO) method, or by replacing such process by a fast decision algorithm that avoids the computation of each block coding cost, aiming to reduce the number of methods used to partition the Coding Tree Units (CTUs) into Coding Units (CUs). This process is even more complex in VVC than in HEVC, given that besides Quadtree (QT) partitions, two more partition types have been added, namely Binary Tree (BT) and Ternary Tree (TT) partitions. To tackle this problem, Na Tang *et al*. leverage the Canny Edge detector to preform early termination if the CU is uniform enough, or select horizontal or vertical partitions, depending on the ratio between the number of horizontal and vertical detected edges [8]. Jing Cui *et al*. base their decision upon the gradients of the CU, to choose what partition type should be applied to the CU [2]. Thomas Amestoy *et al*., take advantage of a number of features fed into a set of Random Forests models trained for each partition depth, to decide whether QT or BT should be applied [1]. Genwei Tang *et al*., on the other hand, propose a Split/No-Split approach where a shape-adaptive Convolutional Neural Network replaces the RDO process and to decide whether a given CU should be or not further split [7].

In this paper, we propose an off-loop early termination method, that leverages three Extremely Randomised Trees (ERT) models to predict the maximum partition depth per partition type (QT, BT and TT), that can be achieved in a given CTU. Once the maximum partition depth for a given partition type is achieved, no further partition of the same type is performed. The remainder of the paper is organised as follows: section 2 presents a detailed description of the proposed method, section 3 presents the achieved results and, finally, some conclusions are drawn in section 4.

Table 1: Summary of the used features.

ID	Feature
1	Latitude of the centre point of the CTU
2	Secant of the latitude of the centre point of the CTU
3	Spatial Information
4	Std. Dev. of Sobel filtered CTU along x
5	Std. Dev. of Sobel filtered CTU along y
6	Std. Dev. of Sobel filtered bottom left fourth of CTU, along y

2 Proposed Method

The proposed method extracts off-loop features from a given CTU, feeds them into tree ERT models (one for each type of partition, *i.e*. QT, BT, and TT), that predicts the maximum depth (per partition type). This limits the partition depth that is going to be tested by the RDO process. For example, if the models predict that the maximum depth of the QT, BT, and TT partitions are 2, 2, and 2, respectively, it means that only two depths of each partition type will to be tested, greatly reducing the total number of hypothesis to be tested using RDO, and thus reducing the overall complexity of the encoding process.

It is worthwhile to notice that not all CUs resulting from the RDO process present the maximum estimated complexity, since larger CUs may be more suitable to certain regions of the CTU. However, in order to limit the impact of this early termination method in the coding efficiency, only the maximum depths are estimated. In other words, if the predictive models had 100% accuracy, this method would have had absolutely no impact on the coding efficiency. Therefore, all coding efficiency losses are directly caused by the models mis-classification.

2.1 Features

Initially, a set of 56 features was extracted from the CTUs of the training dataset. Then, ERT models were used to preform Recursive Feature Elimination (RFE).

Features 1 and 2 from Table 1 take advantage of the geometric characteristics of the Equirectangular Projection (ERP) [6]. As noticed in [4], most of the coding complexity related to the ERP format is clustered near the equator, while regions near the poles tend to require lower complexity. Furthermore, it can be demonstrated that the ERP distortion that originates regions of lower complexities has a direct relationship with the $secant(l)$ of the latitude (l). Therefore, Feature 1 discriminates the vertical position of a given CTU, while Feature 2 is related to the distortion that

spawns the low complexity regions. Feature 3 is the Spatial Information [9], while Features 4 and 5 discriminate between the horizontal and vertical directions, respectively. Finally, to capture finer detail, the CTUs were split into four smaller squares of 64 by 64 pixels, and the same spatial features were computed to each of these squares. The standard deviation of the bottom left square along the y direction was deemed relevant to predict the maximum CTU partition depth, by the RFE process.

2.2 ERT Model

The advantages of ERT over Random Forests, arise from the method used to choose attributes and cut-points while cutting a tree node. In Random Forest this is done by finding the local optimum cut-point for each feature, using a metric such as Information Gain. In ERT, a set of cut-points is randomly generated for each feature. Then, the cut-point from the set that yields the best accuracy is selected. Furthermore, in ERT each decision tree is trained over the entire training dataset.

The three ERT models were trained to predict the maximum partition depth for each of the partition schemes (QT, BT, and TT). To achieve this, a training/testing dataset was generated by encoding the 10 ERP sequences recommended by [3], in all intra configuration and $QP = 22$, and then registering the maximum depth achieved by each partition type for all CTUs in the frame. Furthermore, a set of 6 features mentioned in Section 2.1 was extracted for each CTU.

Finally, the ERT models were trained and tested using Cross-Validation, such that the models were trained using data from 9 sequences and then tested against the remaining one. Using this methodology, the QT model achieved an Average Accuracy on the test dataset of $71 \pm 6\%$, the BT model $67 \pm 12\%$, and the TT model $92 \pm 7\%$. Leveraging the 0-1 loss metric, we can approximate the bias and the variance of each of the 3 models. The QT model presents an estimated bias of 29% and an estimated variance of 6%, the BT model bias of 33% and variance of 12%, and the TT model bias of 8% and variance of 7%. This shows that the TT model presents low levels of both bias and variance, as desired, indicating that this model present neither over nor underfitting. Regarding the QT and BT models, a relatively low variance and higher bias is presented, indicating some level of underfitting.

After the three models have been trained, they were implemented within the VVC encoder and the respective functions are called each time a new CTU is encoded. Then, the partition depth limits are updated according to the prediction of the models. Some constrains were implemented in the encoder, so that no partition depths above the predicted maximum are evaluated by the RDO process. It is worthwhile to note that this off-loop approach has the advantaged that the models are required to run only once per CTU, resulting in negligible complexity overhead. In-loop approaches often have to compensate the introduced overhead, since their functions are typically called several times during the encoding of a single CTU.

3 Results

The proposed method was evaluated by measuring the processing time required to encode each of the 10 sequences, and then comparing the time and coding efficiency with the same sequences encoded using the standard VVC reference software (VTM 8.0). All sequences have a 4432 by 2216 resolution, and were encoded using all intra configuration, next profile, and a set of 4 QPs (22, 27, 32, and 37), in order to compute Bjontegaard Delta Rate (BD-Rate). This metric was used to evaluate the coding efficiency, while the complexity was evaluated by computing the average across the 4 QPs of the difference between the encoding time using the proposed method and using the reference VVC, normalised to the encoding time of the latter.

Table 2 shows these results for all 10 sequences. In all cases, the proposed method presents significantly reduced complexity, when compared to the unaltered implementation of VVC, with a negligible loss of coding efficiency, that is less than 1% for 9 out of the 10 cases. In fact, the proposed method presents on average 31.35% complexity reduction, with an average increase in bitrate for a given visual quality of about 0.42%.

Moreover, if one divides the average complexity reduction by the BD-Rate, to determine the percentage of complexity reduction that a given method can achieve per each 1% of BD-Rate loss, we can conclude that the proposed method achieves a ratio of 74.30, outperforming other state of the art methods, such as [8] (23.39), [7] (33.75), [2] (50.00), and [1] (52.63).

Table 2: Results for the proposed method.

Sequence	BD-Rate (%)	Avg. Complex. Reduction (%)
Harbor	0.79	−26.00
KiteFlite	0.42	−28.46
Balboa	0.36	−31.43
BranCastle	0.24	−35.71
Broadway	0.36	−31.23
Landing2	0.30	−35.88
SkateBoardInLot	0.96	−36.45
ChailiftRide	1.06	−35.92
Trolley	0.42	−31.27
Gaslamp	0.72	−25.69
Average	0.42	−31.35

4 Conclusions

In this paper we propose a novel algorithm, that leverages 3 ERT models to predict the maximum partition depth of each partition type (QT, BT, and TT) for every CTU in intra frames of 360° video sequences in ERP format. The prediction is used in a modified version of the VVC, to limit the RDO process to depths smaller than the maximum partition depths predicted by the models. The proposed method achieves an average complexity reduction of 31.35%, with a negligible average coding efficiency loss of just 0.42%. Additionally, the proposed method is able to outperform other state of the art low complexity solutions, such as [7, 8].

Acknowledgements

This work was supported by Programa Operacional Regional do Centro, project ARoundVision CENTRO-01-0145-FEDER-030652 and by FCT/MCTES through national funds and when applicable co-funded EU funds under the project UIDB/EEA/50008/2020, Portugal.

References

[1] T. Amestoy, A. Mercat, W. Hamidouche, C. Bergeron, and D. Menard. Random forest oriented fast qtbt frame partitioning. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1837–1841, 2019.

[2] J. Cui, T. Zhang, C. Gu, X. Zhang, and S. Ma. Gradient-based early termination of cu partition in vvc intra coding. In *2020 Data Compression Conference (DCC)*, pages 103–112, 2020.

[3] P. Hanhart, J. Boyce, K. Choi, and J.-L. Lin. L1012: JVET common test conditions and evaluation procedures for 360° video. Technical report, Joint Video Experts Team (JVET), 12th Meeting: Macau, CH, October 2018.

[4] B. Ray, J. Jung, and M. Larabi. A low-complexity video encoder for equirectangular projected 360 video content. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1723–1727, April 2018. doi: 10.1109/ICASSP.2018.8462368.

[5] R. Skupin, Y. Sanchez, Y. . Wang, M. M. Hannuksela, J. Boyce, and M. Wien. Standardization status of 360 degree video coding and delivery. In *IEEE Visual Communications and Image Processing (VCIP)*, pages 1–4, December 2017. doi: 10.1109/VCIP.2017.8305083.

[6] John P. Snyder. Map projections: A working manual. Technical report, U.S. Government Printing Office, 1987.

[7] G. Tang, M. Jing, X. Zeng, and Y. Fan. Adaptive cu split decision with pooling-variable cnn for vvc intra encoding. In *2019 IEEE Visual Communications and Image Processing (VCIP)*, pages 1–4, 2019.

[8] N. Tang, J. Cao, F. Liang, J. Wang, H. Liu, X. Wang, and X. Du. Fast ctu partition decision algorithm for vvc intra and inter coding. In *2019 IEEE Asia Pacific Conference on Circuits and Systems (APCCAS)*, pages 361–364, 2019.

[9] H. Yu and S. Winkler. Image complexity and spatial information. In *Fifth International Workshop on Quality of Multimedia Experience (QoMEX)*, pages 12–17, July 2013. doi: 10.1109/QoMEX.2013.6603194.

[10] Fan Zhang, Angeliki V. Katsenou, Mariana Afonso, Goce Dimitrov, and David R. Bull. Comparing vvc, hevc and av1 using objective and subjective assessments, 2020.

Deep Learning Algorithms for Tissue Identification in Hysteroscopies

Ana Martins[1]
up201405006@fc.up.pt
Francesco Renna[1]
frarenna@dcc.fc.up.pt
Mihaela Gotseva[2]
mihaela.gotseva@gmail.com
Hélder Ferreira[2]
helferreira78@gmail.com
Miguel Coimbra[3]
mcoimbra@dcc.fc.up.pt

[1] Instituto de Telecomumicações
Faculdade de Ciências da Universidade do Porto
Porto, PT
[2] Centro Hospitalar Universitário do Porto
Hospital de Santo António
Porto, PT
[3] INESC TEC
Faculdade de Ciências da Universidade do Porto
Porto, PT

Abstract

In this work, we present a comparison of different deep learning methods to classify images of uterine tissue collected from hysteroscopy exams. The considered solutions are based on the use of different convolutional neural networks and transfer learning strategies and they are applied to two distinct classification problems: i) fully automatic classification of hysteroscopy images and ii) semi-automatic classification of pre-selected portions of hysteroscopy images.

The obtained results testify the potential limitations of deep learning approaches in the presence of very limited training data in the detection of uterine polyps from normal endometrial tissue, where a maximum accuracy of 74% has been achieved. On the other hand, when applied to a semi-automatic task where significant portions of the images are pre-selected, the considered deep learning solutions achieve accuracy values above 92%, also in the presence of a reduced amount of training data.

1 Introduction

Hysteroscopy is a routine gynaecological procedure, which involves insertion of a small camera transvaginally into the uterine cavity in order to identify abnormalities, and in many cases treat them at the same time. As with any surgical procedure, there is a risk of complications, which is overall very low, however in some cases they can have serious long-term consequences. One of the most relevant complications is uterine perforation (UP). The reported incidence of UPs varies from country to country and is reported between 0.12 to 3% in Germany [2], Holland [3], and France [1]. The reason UPs are a concern is that in rare cases they can lead to major haemorrhage, which can require a life-saving hysterectomy. In other cases, UPs can be associated with injury to the bowel, bladder and ureters which often require additional surgical procedures and long-term treatment. In the context of pregnancy, UPs can lead to uterine dehiscence during pregnancy or delivery, which can be life-threatening for the mother and child. Another very rare long-term complication is the formation of fistulas between the abdomen and the uterus.

These rare, but potentially severe complications underline the need for creating computer assisted decision (CAD) systems for hysteroscopy, able to actively recognize the different kinds of tissues explored during the exam in order to further increase the safety of the procedure. A first step in this direction is represented by the development of a classification algorithm able to differentiate different types of uterine tissues from images collected during hysteroscopy.

Although, to the authors knowledge, there are no works in the literature that specifically addressed the problem of classifying images collected during hysteroscopy exams, deep convolutional neural networks (CNNs) are currently regarded as the state-of-the-art for several related biomedical image classification applications. For example, a study done for the classification of endoscopy images of small intestine tissue based on CNNs achieved higher classification sensitivity and shorter reading times than a conventional analysis done by gastroenterologists [5]. Similarly, deep neural networks have been shown to outperform doctors in the accurate differentiation of tiny colorectal polyps [4].

In this paper, we consider two classification tasks on hysteroscopy images that aim to discriminate between normal endometrial tissue and endometrial polyps (Figures 1 and 2). The first task consists in a fully automatic classification of hysteroscopy images, whereas the second task depicts a semi-automatic scenario where pre-selected cropped images are classified via deep CNNs.

2 Methodology

2.1 Materials

A total of 270 images of size 720×576 were collected from 25 patients during hysteroscopy exams perfomed in an outpatient clinic (OC) scenario. In addition, further 230 images were extracted from 11 videos of resolution 1440×1080 recorded during hysteroscopy exams performed under general anaesthetic (GA) in the operating room.

The images in the obtained dataset were divided into two classes by an experienced gynaecologist: normal endometrial tissue (Figure 1) and endometrial polyps (Figure 2). The first class contained 140 images of 13 patients from OC hysteroscopies plus 110 images extracted from hysteroscopy videos of 8 patients. Moreover, 130 images of 12 OC patients plus 120 video frames from 7 GA hysteroscopy patients were included in the second class (Table 1).

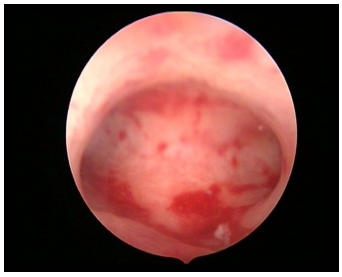


Figure 1: Example of image of normal endometrium tissue.

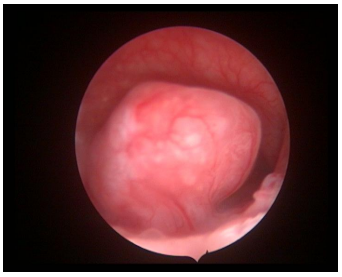


Figure 2: Example of image of endometrial polyp .

The dataset for the semi-automatic classification task was generated from the previous dataset (500 images from 40 patients) by cropping four different significant portions from each image.

Tissues	Nº of images	Cropped images	Nº of patients
Normal endometrial	140+110	1000	13+8
Endometrial polyp	130+120	1000	12+7

Table 1: Division of images into normal endometrial tissue and endometrial polyp classes

2.2 CNN architectures and training

In this work, two different convolutional neural network architectures were considered: VGG-16 and ResNet-50, as they demonstrate excellent performance in a variety of related biomedical image classification tasks. In addition, the sets of weights of these architectures trained over the ImageNet dataset are publicly available.

VGG-16 and ResNet-50, as they demonstrate excellent performance in a variety of related biomedical image classification tasks.

Thus, for each network architecture, VGG-16 and Resnet-50, three different transfer learning schemes were considered: i) combination of feature extraction and fully connected layers (FE+FC), ii) combination of

feature extraction and suppor vector machines (FE+SVM), and iii) fine tuning of convolutional and fully connected layers (FT+FC) , where in all three configurations, the networks were pre-trained over the ImageNet dataset.

Feature extraction consists of freezing the convolutional base of the pre-trained model to prevent the weights of these layers from being updated during training. On the other hand, the fully connected layers of the network are trained from scratch with the data of the considered task, to allow adaptation of the classification to the data set and the analyzed classes. When combining feature extraction with a support vector machines (SVM), the features obtained from the convolutional layers of the pre-trained networks are used as input of an SVM which is trained over the data of the considered task.

In the case of fine adjustment, only the four layers of the convolutional base are frozen for both VGG-16 and ResNet-50. The remaining convolutional layers and fully connected layers are fine tuned using the data of the considered task in order to extract features more related to the particular classification task and to allow better adaptation of the classifier. Note that re-training some of the convolutional layers with the dataset of the specific task considered allows a greater adaptation of the network for the classification objective, but reduces the robustness against overfitting, given the greater number of parameters trained with the small size dataset.

In order to better cope with the reduced size of the available training dataset, data augmentation is applied to all the training configurations. In particular, for each of the training images, 5 different transformations were considered including rotations, mirroring, zooming, and brightness level adjustment.

All networks were trained for 50 epochs, using the Adam optimizer with learning rate 0.0001 and mini-batch size of 32. Additionally, a dropout of 0.4 was used in two layers for each network, between the fully connected layers.

3 Results

In this section, we report the classification results obtained with the different CNN-based setups described in Section 2.2 for the fully-automatic and semi-automatic classification of endometrial images. The classification performance is evaluated using the following metrics: accuracy, precision, recall, and F1-score.

For both classifications task, the images in the dataset were randomly divided into 80% training images and 20% test images, guaranteeing that images from patients in the test set could not be included in the training set. The classification results for this task are reported in Table 2.

	VGG-16			ResNet-50		
Valores	**FE+FC**	**FT+FC**	**FE+SVM**	**FE+FC**	**FT+FC**	**FE+SVM**
Accuracy	0.67	0.54	0.64	0.70	0.74	0.70
Precision	0.69	0.52	0.60	0.70	0.67	0.64
Recall	0.62	0.90	0.86	0.70	0.94	0.92
F1-score	0.65	0.66	0.70	0.70	0.78	0.75

Table 2: Comparison of different transfer learning techniques applied to the VGG-16 and ResNet-50 architectures for the fully automatic classification.

It can be observed that the classification performance is, in general, not very satisfactory, even if a slight advantage is obtained when using the ResNet-50 architecture. The poor performance registered is mainly caused by the lack of a larger training set, thus leading to significant overfitting, and by the presence of specific features in the images that can lead to misclassification. In particular, several errors are observed in the classification of images of normal endometrial tissue, since a significant portion of them contains the channels of the fallopian tubes (Figure 3), which are often confounded with the presence of polyps. On the other hand, the proposed algorithms often fail in detecting small polyps from images (Figure 4).

Table 3 contains the results obtained when applying the CNN-based methods described in Section 2.2 to the semi-automatic task of classifying pre-selected cropped images from the original dataset. In this case the proposed, architectures are able to achieve significantly better performance, thus guaranteeing reliable discrimination between endometrial polyps and normal tissue.

Figure 3: Example of image from normal endometrium with fallopian tube channel.

Figure 4: Example of image with the presence of a small polyp.

	VGG-16			ResNet-50		
Valores	**FE+FC**	**FT+FC**	**FE+SVM**	**FE+FC**	**FT+FC**	**FE+SVM**
Accuracy	0.96	0.93	0.92	0.95	0.95	0.96
Precision	0.97	0.89	0.89	0.92	0.92	0.94
Recall	0.94	0.97	0.96	0.99	0.98	0.99
F1-score	0.96	0.93	0.93	0.96	0.95	0.97

Table 3: Comparison of techniques in the transfer learning application to the VGG-16 and ResNet-50 architecture for the semi-automatic classification.

The data considered for this tasks are portions of the original images, which may facilitate the training of the network due to the 4x increase of the dataset size. This has allowed the network to extract the necessary characteristics in order to distinguish the classes. Moreover, the considered cropped images represent lower-dimensional data with reduced variability, thus simplifying the corresponding classification task.

4 Conclusion

The problem of classifying images obtained from an hysteroscopy exam using CNN-based classifier was considered. Different network architectures and transfer learning techniques were tested to discriminate normal endometrial tissue images from endometrial polyps.

When considering a fully automatic classification of hysteroscopy images, the use of fine tuning on a ResNet-50 architecture pre-trained over the ImageNet dataset is shown to provide interesting classification results even in the presence of a strictly reduced training dataset.

On the other hand, classification of pre-selected portions cropped from the original images is shown to be reliably performed even with such a small training dataset, due to the reduced variability of the considered samples.

Acknowledgments

This work is funded by FCT/MCTES through national funds and when applicable co-funded EU funds under the project UIDB/50008/2020.

References

[1] Aubert Agostini et al. Risk of uterine perforation during hysteroscopic surgery. *The Journal of the American Association of Gynecologic Laparoscopists*, 9:264–7, 08 2002. doi: 10.1016/S1074-3804(05)60401-X.

[2] Burkhard Aydeniz et al. A multicenter survey of complications associated with 21 676 operative hysteroscopies. *European journal of obstetrics, gynecology, and reproductive biology*, 104:160–4, 10 2002. doi: 10.1016/S0301-2115(02)00106-9.

[3] Frank William Jansen et al. Complications of hysteroscopy: a prospective, multicenter study. *Obstetrics and gynecology*, 96 2:266–70, 2000.

[4] Peng-Jen Chen et al. Accurate classification of diminutive colorectal polyps using computer-aided analysis. *Gastroenterology*, 154, 10 2017. doi: 10.1053/j.gastro.2017.10.010.

[5] Zhen Ding et al. Gastroenterologist-level identification of small bowel diseases and normal variants by capsule endoscopy using a deep-learning model. *Gastroenterology*, 2019.

Ship Segmentation in Aerial Images for Maritime Surveillance

Carlos Pires[1]
c.david.pires@tecnico.ulisboa.pt
Alexandre Bernardino[1]
alex@isr.tecnico.ulisboa.pt
Bruno Damas[1]
bdamas@isr.tecnico.ulisboa.pt

[1] Institute for Systems and Robotics, ISR
Instituto Superior Tecnico, IST
Universidade de Lisboa,
Lisbon, PT

Abstract

In this paper, we study and implement a method to detect ships during maritime surveillance missions. We implement a cascade model with a detection part followed by a segmentation stage. We use two convolutional neural networks, one for each section. With detection, we select the most likely regions to contain a ship, and after we segment those regions to identify the targets. We train the model with maritime datasets, and then we test it on the Airbus Ship detection challenge. The cascade model is capable of real-time ship segmentation, achieves a score of 0,82 in the challenge, and processes one image in 0,1 seconds.

1 Introduction

Maritime surveillance is a need for a country with a coast to prevent, discourage, and punish catastrophic and illegal events. Therefore, it is necessary to control all maritime activities. According to [1], Portugal has a coast with 943 kilometers and an exclusive economic zone (EEZ) with almost two million square kilometers divided into three regions. So, due to the amount of area, it is crucial to develop efficient and cost-effective tools. Unmanned Aerial Vehicles (UAVs) are flexible and extensible systems that can incorporate a vast number of sensors [2]. With them, we release a considerable amount of human resources. Image segmentation breaks the image down into meaningful regions, and we can separate a ship from the rest of the frame and estimate the ship size and route. Then, it is possible to compare the data with the automatic identification system (AIS). Plus, after applying segmentation, we can use techniques on the pixel level to improve ship detection performance. Object segmentation is a challenging task, especially when we are using images captured by UAVs because it is affected by factors like scale, perspective, and illumination variations. Images with glare and waves may confuse the model, which makes the segmentation task harder. The goal of this work is to implement an automatic real-time maritime surveillance system using drones with onboard cameras.

The main contribution is the development of a two stages system to perform ship segmentation. We present a cascade model with detection and segmentation, which makes the ship identification faster. The algorithm has two different stages: first, we use a fast detection network to search for possible ship locations regions, then we pass these regions through a segmentation network, thus narrowing the image region where segmentation is performed, which significantly improves the overall image processing time.

2 Related work

Since 2012, when A. Krizhevsky *et al*. [3] trained a convolutional neural network (CNN) on ImageNet and achieved outstanding results, the use of deep learning methods for detection and segmentation has increased exponentially. They have shown that with more layers (more depth), networks exhibit significant performance improvement. In detection algorithms, we try to generate a bounding box (BB) around the detected objects. In [4], R-CNN, Region with CNN features, takes an image as input and identifies where the primary target is. [5] presents a different approach to perform object detection with a network called YOLO - You Only Look Once. They treat object detection as a regression problem to separate BBs and to associate class probabilities. YOLO applies a single CNN to the full image, divides the image into regions, predicts BBs, and a score for each one.

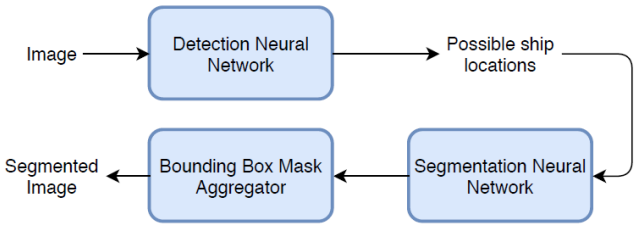


Figure 1: Segmentation system global architecture.

To successfully train a deep neural network many training samples are required. [6] presents an architecture and a training strategy with effective use of data augmentation to use the available training samples efficiently. The classification network proposed, U-Net, performs image segmentation by class and predicts a mask, which separates all the objects. This network has a U structure, and on the left side is done a contracting path to capture context, encoder. The right side has an expanding part that enables precise localization.

The authors of [7] use recorded detect data to help to detect a vessel. They investigate how temporal features could improve ship detection in video sequences captured by small aircraft. So, they build a convolutional Long short-term memory to learn those features and increase ship detection. [8] proposes a ship detection method in challenging surveillance conditions. They track vessels with a correlation filter complemented with image segmentation. To compensate for drifts in the correlation filter, they apply blob analysis to re-center the target in the tracking window. [9] presents a robust detector for aerial images. They modify a CNN and create tracks with the successive detections to predict the position of the objects in future frames.

3 Methodology

Our algorithm contains two phases: first, it applies image detection to identify possible locations of the ship, and then, these regions are segmented to check if there are ships. With these two parts, we intended to turn the segmentation task faster and better because we discard a considerable amount of background. So, now we are focused on segmenting particular regions of the image with a high probability of containing a ship. Then, the possibility of over-segmentation and miss segmentation will decrease. Figure 1 shows the system global architecture.

Our approach uses a detection network before segmentation to delete a considerable part of the image background. So, the cascade model with detection and segmentation has to be quicker than full image segmentation. We will use the YOLO network since it can process images in milliseconds, and it is computationally efficient. We implemented the YOLO-tiny version, which has 24 layers, and most are convolutional and max-pooling layers. In YOLO, we can adjust the detection threshold: when we set it to a specific value, the network only displays objects detected with a confidence score above that value. So, we can build a rigid detection model which only detects objects with a high detection probability or a permeable model that detects objects with a low confidence score. In the first case, we could miss a vessel in the image, and in the second case, the network can identify ocean regions as a ship. In this first stage, it is necessary to have a high recall. Even if we get some false positives, there is no problem because, in the segmentation stage, we will filter them. Therefore, regions with waves or sun glare are mostly like to be in the region proposals. These areas can be easily confused as a ship.

Dataset	Recall	True Positive	False Negative	False Positive
kaggle	0,9	63	7	504
Seagull	0,86	115	19	1233

Table 1: Detection network results with a threshold=0,001.

Label	Dataset	NºBB	IoU
From dataset	Kaggle	382	0,94
Our	Kaggle	386	0,89
Our	Seagull	494	0,91

Table 2: Segmentation results with encoder densenet121.

Next, we need to identify the ship using segmentation. Thus, we process an image to identify a class for each image pixel. In our case, there are two labels: background and ship. Following that, we find a cluster that represents the vessel, and we can get the target size and shape. We use the U-Net to perform semantic segmentation because it is efficient and produces good results with few training samples. In image segmentation, we have to convert the feature map into a vector to classify the pixels and then reconstruct the image from this vector. U-net uses the same feature maps that are used for the contraction to expand the vector to a segmented frame. Then, we preserve the structural integrity of the image and decrease the output distortion. Networks like U-net have an encoder-decoder architecture. For the encoder, we will use different architectures and study how it affects performance. Sometimes we have multiple BBs per image, and we need to group them to reconstruct the full segmented mask. We use the BB mask aggregator module.

4 Experiments

Since we are using deep learning methods, we need to train both networks with samples. So, we used RGB images from the Seagull and Kaggle dataset, [10] and [11], respectively. Once we have two networks, we divided the training phase into two steps. First, we trained the YOLO network with images from both datasets, 45000 each. Then, we passed a set of 20000 images through YOLO, and we used the BBs results to train the U-net. For both networks, we split the set in 70%-30% for training and validation. Plus, for segmentation, we tried multiple encoders, like ResNet, Densenet121, and Inceptionresnetv2, all pre-trained on ImageNet. Additionally, we tested various loss functions, focal, dice, and cross-entropy losses, also varying the batch size.

We trained and tested both networks using a Keras implementation with Tensorflow, running on an Nvidia GTX1070 Ti. As a test set, we chose images from both datasets, 575 frames. Since in the Seagull dataset, the label is a BBs, we had to segment each image manually. Furthermore, we tested our cascade model in the Airbus Ship detection challenge [11]. We submitted the segmentation results for a set of images, then based on the f2-score and IoU metrics, we got a Kaggle score.

5 Results

To evaluate our method, we tested both networks. For the detection stage, we searched for the best threshold value, with th=[0,00001; 0.5]. Table 1 shows the accuracy and recall for the best threshold, th=0,001. Despite a high number of BBs provided by YOLO, we remove between 93-95% of the background image, which turns the next stage quicker. To evaluate the segmentation task, we passed through the U-net the BBs from the detection network. Table 2 shows the IoU result for three test sets with densenet121 as an encoder, batch size of 16, and a combination of focal and weighted dice losses. Plus, we submitted our solution to the Kaggle challenge, and we compared it with full image segmentation. Table 3 displays the difference between the two methods, and the cascade model has a better score in less time. According to [11], the best method achieved a score of 0.85, and our method got a lower score, 0,82. However, we decreased the processing time per image by reducing the search space in the segmentation stage. The proposed model provides real-time ship segmentation, by achieving an average processing rate of 10 images second.

Method	Kaggle score	Time/image [s]
Full image segmentation	0,71	1,47
Detection + Segmentation	0,82	0,1

Table 3: Airbus kaggle challenge results.

6 Conclusions

This paper presents a contribution to performing fast and accurate maritime ship segmentation. This method has two main stages, in the first part, we extract possible ship location regions, and then we segment these regions to identify the ship. The initial stage allows removing unnecessary parts of the image to identify the target, which speeds up the segmentation part. In aerial images, a ship is a small target, and it is easily confused with the background. With this cascade model, we decrease detection failures and improve the segmentation mask. We tested our method on the Airbus Ship detection challenge, and we achieved a score of 0,82. The best results got a score of 0,85 in the challenge. However, our cascade model is capable of real-time detection since, on average, an image is processed in 0,1 seconds.

Acknowledgements

This work was supported by FCT with the LARSyS - FCT Project UIDB/50009/2020 and project VOAMAIS (02/SAICT/2017/31172).

References

[1] Direção-Geral de Recursos Naturais, Segurança e Serviços Marítimos. Zonas marítimas sob soberania e ou jurisdição portuguesa. https://www.dgrm.mm.gov.pt/am-ec-zonas-maritimas-sob-jurisdicao-ou-soberania-nacional, n.d. Accessed: 2020-09-7.

[2] Gonzalo Pajares. Overview and current status of remote sensing applications based on unmanned aerial vehicles (uavs). *Photogrammetric Engineering and Remote Sensing*, 81(4):281 – 329, 2015.

[3] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012.

[4] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.

[5] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *CoRR*, abs/1804.02767, 2018.

[6] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, 2015.

[7] G. Cruz and A. Bernardino. Learning temporal features for detection on maritime airborne video sequences using convolutional lstm. *IEEE Transactions on Geoscience and Remote Sensing*, 57(9):6565–6576, 2019.

[8] J. Matos, A. Bernardino, and R. Ribeiro. Robust tracking of vessels in oceanographic airborne images. In *OCEANS 2016 MTS/IEEE Monterey*, pages 1–10, 2016.

[9] G. Cruz and A. Bernardino. Evaluating aerial vessel detector in multiple maritime surveillance scenarios. In *OCEANS 2017 - Anchorage*, pages 1–9, 2017.

[10] R. Ribeiro, G. Cruz, J. Matos, and A. Bernardino. A data set for airborne maritime surveillance environments. *IEEE Transactions on Circuits and Systems for Video Technology*, 29(9):2720–2732, 2019.

[11] Airbus. Airbus ship detection challenge. `https://www.kaggle.com/c/airbus-ship-detection/overview`. Accessed: 2020-09-5.

Comparison and Evaluation of Information-based Measures in Images

Jorge Miguel Silva
jorge.miguel.ferreira.silva@ua.pt
IEETA,
University of Aveiro,
Aveiro, Portugal

Diogo Pratas
pratas@ua.pt
Department of Virology,
University of Helsinki,
Helsinki, Finland

Sérgio Matos
aleixomatos@ua.pt
DETI,
University of Aveiro,
Aveiro, Portugal

Abstract

Lossless data compressors and small Turing machines can approximate the quantity of information present in a digital object. In this paper, we describe and compare these approaches of measuring unsupervised probabilistic and algorithmic information on images (2D) with different characteristics. We use the Normalized Compression (NC) employing the data compression PAQ8 and compare it with the Block Decomposition Method (BDM) and show some advantages and limitations of both measures.

1 Introduction

There are several approaches to quantify the amount of information. Kolmogorov described three, namely combinatorial, probabilistic, and algorithmic [4]. While the Kolmogorov complexity is non-computable, it can be approximated with programs for such purpose, such as data compressors, using probabilistic and algorithmic schemes. Practical applications to approximate the Kolmogorov complexity for multiple dimensional digital objects have been developed using Turing machines [6, 7] and data compressors [3]. Recently, Zenil *et al.* have shown that this methodology has a closer connection to algorithmic information than other measures based on statistical regularities [7], namely fast lossless compression methods, for sources that follow algorithmic schemes. One of the applications of information theory is to measure image information. Herein, we define an image's quantity of information as the smallest number of bits required by a model to represent an image losslessly. To perform this task, the model searches for unknown patterns of similarity between sub-regions of the image and uses this information to create this compressed representation of the image, relying exclusively on the two-dimensional pixels' patterns without using exogenous information. In this paper, we describe and compare solutions for unsupervised measures of probabilistic and algorithmic information in images (2D) of different datasets. We use the Normalized Compression (NC) employing PAQ8 data compression tool and compare it with the Block Decomposition Method (BDM) [7], and the inherent Coding Theorem Method (CTM) measures [2]. The BDM is an information-based measure that uses small Turing machines to approximate the algorithmic information, approximating to the Shannon entropy as a fallback mechanism.

2 Methods

In this section, we describe the Normalized Compression (NC) and two Block Decomposition Method (BDM) normalizations.

Normalized Compression (NC)

An efficient compressor, $C(x)$, gives a possible approximation for the Kolmogorov complexity ($K(x)$), where $K(x) < C(x) \leq |x|$ ($|x|$ is the length of string x in the appropriate scale). Usually, an efficient data compressor is a program that approximates both probabilistic and algorithmic sources. Although the algorithmic nature may be more complex to model, data compressors may have embedded sub-programs to handle this nature. For a definition of safe approximation, see [1]. The normalized version, known as the Normalized Compression (NC), is defined by $\mathrm{NC}(x) = \frac{C(x)}{|x|\log_2|\mathrm{A}|} = \frac{C(x)}{|x|}$, where x is a string, $C(x)$ is the compressed size of x in bits, $|\mathrm{A}|$ the number of different elements in x (size of the alphabet) and $|x|$ the length of x. Since we consider a binary matrix of each image, $|A| = 2, \log_2 2 = 1$. Given the normalization, the NC enables to compare the information contained in the strings independently from their sizes [5]. If the compressor is efficient, then the compressor is able to approximate the quantity of probabilistic-algorithmic information in data.

Normalized Block Decomposition Method (NBDM)

Another possible approximation to the Kolmogorov complexity is given by the use of small Turing machines, where these small computer programs approximate the components of a broader representation. The Block Decomposition Method (BDM) extends the power of a CTM, approximating local estimations of algorithmic information based on the Solomonoff-Levin's algorithmic probability theory. In practice, it approximates the algorithmic information and, when it loses accuracy, it performs like Shannon entropy. Since in this article we intend to perform a direct comparison of both measures, we first considered the normalization of the BDM (NBDM_1), given by the number of elements (length) of the digital object: $\mathrm{NBDM}_1(x) = \frac{BDM(x)}{|x|\log_2|\mathrm{A}|} = \frac{BDM(x)}{|x|}$. However, the normalization of the BDM is usually performed using a minimum complexity object (BDM_{Min}) and a maximum complexity object (BDM_{Max}). A minimum complexity object is filled with only one symbol, like a binary string of only zeros. In contrast, a maximum complexity object is an object that, when decomposed (by a given decomposition algorithm), yields slices that cover the highest CTM values and are repeated only after all possible slices of a given shape have been used once. Using these two objects, the NBDM_2 for a given string can be computed as $\mathrm{NBDM}_2(x) = \frac{BDM(x)-BDM_{Min}}{BDM_{Max}-BDM_{Min}}$, where $BDM(x)$ is the BDM value of that string, BDM_{Min} is the minimum complexity object, and BDM_{Max} is the maximum complexity object. Kolmogorov complexity is invariant only up to a constant factor, which depends on the choice of a description language $K = K' + L$, where K is the total complexity, K' is the description of the object and L is the description of the language. As such, by performing the normalization according to Equation 2, the normalization is aiming to remove the constant factor as $\frac{K-K_{Min}}{K_{Max}-K_{Min}} = \frac{K'+L-K'_{Min}-L}{K'_{Max}+L-K'_{Min}-L} = \frac{K'-K'_{Min}}{K'_{Max}-K'_{Min}}$, where K_{Max} and K_{Min} are the maximum and minimum Kolmogorov complexity objects and K'_{Max} and K'_{Min} are the maximum and minimum Kolmogorov complexity description of the objects.

3 Results and Discussion

In order to compare NC with BDM, we performed three types of tests. Namely, we compared the robustness of both measures according to increasing rates of random pixel changes in paintings, tested their application on different types of images, and made an assessment of the minimal information bounds. In the first test, we assessed the impact of an increasing rate of pixel editions using a pseudo-random uniform distribution and compared both information-based measures. This approach is not identical to image noise, but rather a pure edition of pixels. For the purpose, we selected a painting from three authors (Theodore Gericault, Marc Chagall, and Rene Magritte), making 50 adulterated copies of each painting with increasing edition rate (from 1 to 50%). Finally, we measured the NC (Eq. 2), the NBDM_1 (Eq. 2), and NBDM_2 (Eq. 2) in all the paintings. Figure 1 (A) depicts the values obtained for the NC and BDM. The results show that, when using the same type of normalization, NC is more robust to the increment of pixel edition than NBDM (NBDM_1). On the other hand, whereas NBDM_1 considers the normalization by the length of the input object, NBDM_2 performs a normalization that aims to mimic the removal of the constant factor related to Kolmogorov complexity (see Eq. 2). Since the NBDM_2 normalization does not take into account the constant of the description language, it shows a more robust behavior than

$NBDM_1$, which increases rapidly with the increase of pixel edition. Since NC and $NBDM_1$ have the same type of normalization, we will focus on comparing these normalizations from now on.

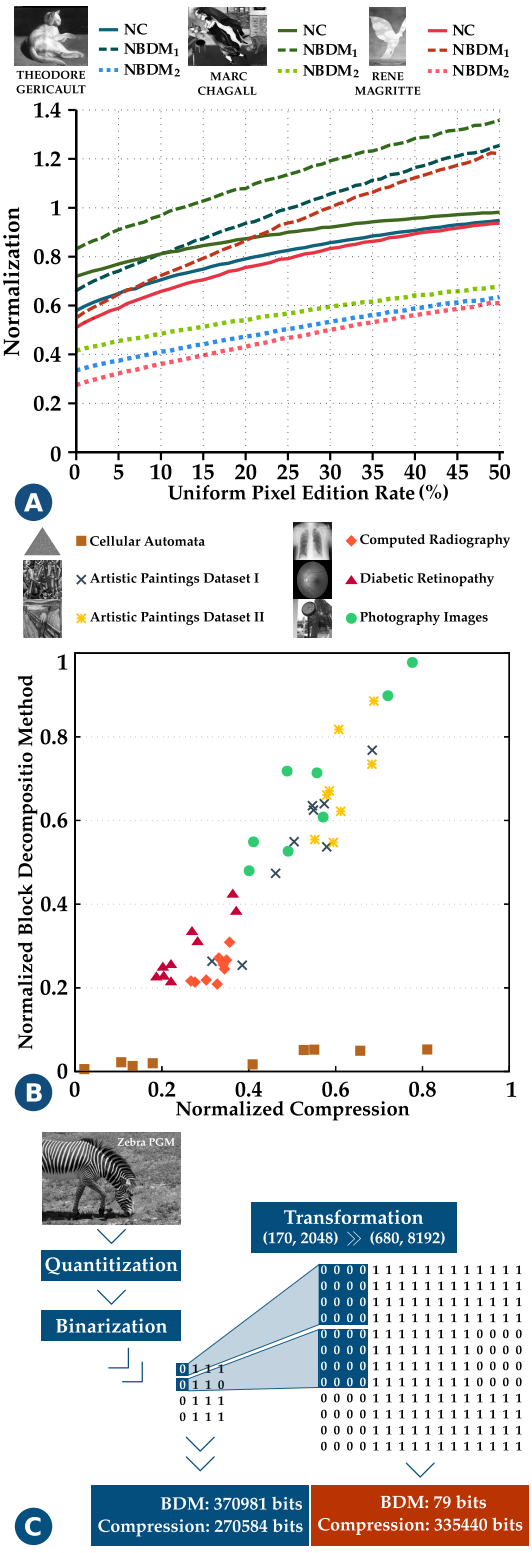


Figure 1: Evaluation of Information-based measures. (**A**) Impact of increasing pseudo-random substitution on information-based measures: NC (approximated using the PAQ8 algorithm) and two BDM normalizations ($NBDM_1$ and $NBDM_2$). (**B**) NC and $NBDM_1$ for different types of images. (**C**) Image transformation pipeline leading to BDM underestimation of the amount of information contained in the transformed object.

In the second test, we applied both measures to six datasets with distinct nature (9 images each) to understand how $NBDM_1$ and NC behave with different types of images. The six datasets were: artistic images from 2 different datasets; cellular automata images; diabetic retinopathy images; chest computed radiography (CR) images and photographic images. The results are depicted in Figure 1 (B). Overall, the majority of the datasets show similar behavior regarding the NC and $NBDM_1$. The exceptions to this are the CR and cellular automata datasets, which exhibit a more algorithmic behavior. The latter dataset is constituted by images created with small programs with simple rules. Whereas the compressor has difficulty compressing this type of images, the BDM can determine their algorithmic nature and thus attribute them with minimal value. This outcome shows the importance of the BDM in the detection of simple output programs embedded into data. In the last test, we selected one of the most complex images identified by the NBDM in the last subsection to test if the BDM could accommodate specific data alterations. This test is depicted in Figure 1 (C). After the binarization process, we performed a super-sample image transformation where each char was amplified to a 4x4 representation. This value was selected since the BDM has the default block size value of 4x4 in 2D structures. After this operation, the BDM was computed for the original and the super-sampled image. While the original image was measured with 370981 bits, the super-sampled image had only 79 bits. This abrupt decrease in the complexity value indicates that the BDM underestimates the amount of information contained in the object. The BDM analyses object information in blocks instead of looking at the whole object. Specifically, blocks analysed by the BDM (default block size value of 4x4 in 2D structures) have the same size as the super-sample image transformation (each char was amplified to a 4x4 representation); therefore, the complexity attributed to each block is approximately zero (since each block is composed of all zeros or ones), and hence the overall value attributed to the complexity of the object will drop dramatically. This analysis shows that BDM is not prepared to deal with the information associated with the choice of the model, unlike the NC. The NC relies on the use of a lossless data compressor, bounded by a maximum information channel capacity.

4 Conclusion

The results show that, when using the same type of normalization, NC is more robust to the increment of pixel edition than NBDM ($NBDM_1$). On the other hand, BDM can determine the algorithmic nature of images created with small programs with simple rules. Whereas the compressor has difficulty compressing this type of image, the BDM can determine their algorithmic nature and attribute them with minimal value. Finally, BDM is not prepared to deal with the information associated with the model's choice, unlike NC. The NC relies on using a lossless data compressor, bounded by a maximum information channel capacity. From these three tests, we can notice some advantages and limitations of both measures. Ranking these measures is not a fair task because they have different characteristics and nature.

References

[1] Peter Bloem, Francisco Mota, Steven de Rooij, Luis Antunes, and Pieter Adriaans. A safe approximation for Kolmogorov complexity. In *International Conference on Algorithmic Learning Theory*, pages 336–350. Springer, 2014.

[2] Jean-Paul Delahaye and Hector Zenil. Numerical evaluation of algorithmic complexity for short strings: A glance into the innermost structure of randomness. *Applied Mathematics and Computation*, 219 (1):63 – 77, 2012. ISSN 0096-3003. doi: https://doi.org/10.1016/j.amc.2011.10.006. Towards a Computational Interpretation of Physical Theories.

[3] Ming Li, Jonathan H. Badger, Xin Chen, Sam Kwong, Paul Kearney, and Haoyong Zhang. An information-based sequence distance and its application to whole mitochondrial genome phylogeny . *Bioinformatics*, 17(2):149–154, 02 2001. ISSN 1367-4803.

[4] Ming Li, Paul Vitányi, et al. *An introduction to Kolmogorov complexity and its applications*, volume 3. Springer, 2008.

[5] Diogo Pratas and Armando J Pinho. On the approximation of the Kolmogorov complexity for DNA sequences. In *Iberian Conference on Pattern Recognition and Image Analysis*, pages 259–266. Springer, 2017.

[6] Fernando Soler-Toscano and Hector Zenil. A computable measure of algorithmic probability by finite approximations with an application to integer sequences. *Complexity*, 2017, 2017.

[7] Hector Zenil, Santiago Hernández-Orozco, Narsis A Kiani, Fernando Soler-Toscano, Antonio Rueda-Toicen, and Jesper Tegnér. A decomposition method for global evaluation of Shannon entropy and local estimations of algorithmic complexity. *Entropy*, 20(8):605, 2018.

Benchmarking bioinspired machine learning algorithms with CSE-CIC-IDS2018 network intrusions dataset

Paulo Ferreira[1]
2180047@my.ipleiria.pt
Mário Antunes[123]
mario.antunes@ipleiria.pt

[1] School of Technology and Management
Polytechnic of Leiria - Portugal
[2] CIIC, Computer Science and Communication Research Centre, ESTG, Polytechnic of Leiria - Portugal
[3] Center for Research in Advanced Computing Systems
INESC-TEC, University of Porto - Portugal

Abstract

This paper aims to evaluate CSE-CIC-IDS2018 network intrusions dataset and benchmark a set of supervised bioinspired machine learning algorithms, namely CLONALG Artificial Immune System, Learning Vector Quantization (LVQ) and Back-Propagation Multi-Layer Perceptron (MLP). The results obtained were also compared with an ensemble strategy based on a majority voting algorithm. The results obtained show the appropriateness of using the dataset to test behaviour based network intrusion detection algorithms and the efficiency of MLP algorithm to detect zero-day attacks, when comparing with CLONALG and LVQ.

1 Introduction

Computer networks security encloses a wide set of technologies to protect the assets and the users operation. Due to its operating mode, Intrusion Detection System (IDS), namely those based on behaviour analysis, are able to detect, with some degree of accuracy and in a timely manner, zero-day attacks and vulnerabilities exploits, to further apply countermeasures. In this paper we intend to evaluate a set of bioinspired algorithms already developed and implemented by Machine Learning (ML) tools. The major contributions can be summarized as follow: i) an open source framework and processing flow, based on WEKA [1], to ingest and process CSE-CIC-IDS2018 dataset; ii) an open source tool to automate the tests carried on with CLONALG [2], LVQ [3] and Backpropagation-MLP [4] classifiers; iii) a comparison between the results obtained individually by each of the bioinspired algorithms with those achieved by an ensemble approach with the same models, using *majority voting* strategy. This paper is organized as follows: Section 2 describes the key concepts for this work. The tests setup is described in section 3, the results are presented in section 4 and further analysed in section 5. Conclusions and future work are described in section 6.

2 Background

IDS can be classified according to the object of analysis (host-based or network-based) and according to the detection method (behaviour-based or signature-based). Behaviour-based IDS aim to overcome the limitations observed on those that are signature-based, namely its inability to detect patterns that are not in a predefined signature database. These systems analyse traffic and try to define a normal network behaviour to further identify deviations that are considered anomalous traffic and, therefore, reported as possible positive examples [5].

Bioinspired ML algorithms are a set of algorithms whose operation is mimicked on systems or mechanisms from the nature or the human body. Some typical applications and analogies are the neural networks, inspired by the functioning of the human brain; the evolutionary and DNA computing, based on theories of evolution that leads to genetic algorithms; the Artificial Immune Systems (AIS), which takes inspiration on the vertebrate immune system, namely its adaptive part [6]. Regarding Artificial Neural Networks (ANN) algorithms, in this work we have used Back-propagation Multi-Layer Perceptron (MLP) [4] and Learning Vector Quantization (LVQ) [3]. From the whole plethora of immune-inspired algorithms [7], the one chosen for this work was CLONal selection ALGorithm (CLONALG) [2].

The tests were carried out with the CSE-CIC-IDS2018 public dataset[1]. Despite being recent, CSE-CIC-IDS2018 dataset is very well organized and is now starting to be widely used by the scientific community to benchmark IDS. It includes a wide range of attacks, executed with different tools, organized in a timeline and mixing both normal and anomalous network packet flows. The traffic was dynamically generated, with the purpose of simulating a corporate network.

Due to the wide variety of attacks and the deluge of data available, we have defined a subset of attacks that could better test the detection of a previously unseen attack. The choice was also based on the diversity and amount of data related to each attack. Table 1 describes the characterization of the attacks used in the experiments carried on in this paper.

Table 1: Network attacks characterization

Date	Time		Type of attack	Software Tool	# flows
	Begin	End			
16/02/2018	10:12	11:08	DoS	SlowHTTPTest	139890
	13:45	14:19	DoS	Hulk	461912
21/02/2018	10:09	10:43	DDoS	LOIC-UDP	1730
	14:05	15:05	DDoS	HOIC	686012

The number of normal traffic flows available at each date is 446772 and 360833 respectively for 16/02/2018 and 21/02/2018.

3 Tests setup

We have carried out four test scenarios, as can be seen on table 2.

Table 2: Test scenarios

Scenario	Training		Testing	
	Date	Traffic	Date	Traffic
1	16/02/2018	Normal+Attack1	16/02/2018	Normal+Attack2
2	16/02/2018	Normal+Attack1	21/02/2018	Normal+Attack1
3	16/02/2018	Normal+Attack2	21/02/2018	Normal+Attack2
4	16/02/2018	Normal+Attacks	21/02/2018	Normal+Attacks

The tests were performed on a subset with 200,000 instances, that is network flows. From that value, 70% (140,000 records) of them constitute the training dataset and the remaining 30% (60,000 records) are part of the testing dataset. The training set records are selected from the training data file and the test set records are selected from the test data file. Each test scenario was then run ten times, with independent data for each iteration, but the same for the three algorithms in each iteration.

Besides the three algorithms mentioned above, we have also considered an *ensemble* of the models generated by the three algorithms, in which the decision strategy is based on the criterion for majority decision, also known as *majority voting*.

The methodology used to run the experiments consists of four main phases: input data ingestion, data preprocessing, data processing and presentation of results (see figure 1).

Figure 1: Methodology

The preprocessing phase deals with issues like removing unnecessary attributes, normalizing data, reducing the number of classes by aggregating every class not being "Benign" as malicious traffic and dealing with missing values by replacing them with the average value for each attribute. These tasks were essentially accomplished through WEKA [1] and Orange [8] *open-source* applications.

[1] https://registry.opendata.aws/cse-cic-ids2018/

The preprocessed dataset is then processed by the algorithms in both training and testing phases. We have used WEKA for that purpose and have also developed an application to automate the tests for any dataset that meets the requirements[2].

4 Results

Tables 3 through 6 show the results obtained for each of the scenarios listed in Table 2. For a given algorithm, the values of each metric correspond to the arithmetic mean of the values obtained for all the ten iterations.

Table 3: Results for scenario 1

Algorithm	TPR	TNR	FPR	FNR	Precision	Recall	Accuracy	F1
CLONALG	0,0306	0,9997	0,0003	0,9694	0,9895	0,0306	0,5071	0,0593
LVQ	0,0306	0,9996	0,0004	0,9694	0,9889	0,0306	0,5071	0,0593
MLP	0,0001	1,0000	0,0000	0,9999	1,0000	0,0001	0,4917	0,0001
Ensemble	0,0306	0,9997	0,0003	0,9694	0,9895	0,0306	0,5071	0,0593

Table 4: Results for Scenario 2

Algorithm	TPR	TNR	FPR	FNR	Precision	Recall	Accuracy	F1
CLONALG	0,0080	0,6537	0,3463	0,9920	0,0337	0,0080	0,6506	0,0103
LVQ	0,7025	0,0031	0,9969	0,2976	0,0034	0,7025	0,0065	0,0067
MLP	0,0000	0,9998	0,0003	1,0000	0,0000	0,0000	0,9950	0,0000
Ensemble	0,0080	0,6537	0,3463	0,9920	0,0337	0,0080	0,6506	0,0103

Table 5: Results Scenario 3

Algorithm	TPR	TNR	FPR	FNR	Precision	Recall	Accuracy	F1
MLP	1,0000	0,9998	0,0002	0,0000	0,9999	1,0000	0,9999	0,9999
CLONALG	1,0000	0,0026	0,9974	0,0000	0,6559	1,0000	0,6562	0,7922
LVQ	1,0000	0,0004	0,9997	0,0000	0,6554	1,0000	0,6554	0,7918
Ensemble	1,0000	0,0026	0,9974	0,0000	0,6559	1,0000	0,6562	0,7922

Table 6: Results for Scenario 4

Algorithm	TPR	TNR	FPR	FNR	Precision	Recall	Accuracy	F1
MLP	0,8977	0,9996	0,0004	0,1023	0,9284	0,8977	0,9327	0,8987
LVQ	1,0000	0,0008	0,9992	0,0000	0,6560	1,0000	0,6561	0,7923
CLONALG	0,9992	0,0030	0,9970	0,0008	0,6564	0,9992	0,6564	0,7923
Ensemble	0,9992	0,0033	0,9968	0,0008	0,6564	0,9992	0,6565	0,7923

5 Results Analysis

The purpose of the tests was to simulate the detection of a *zero-day* attack, by using the CSE-CIC-IDS2018 dataset. It is appropriate to mention that a network attack is essentially an anomaly to the normal network traffic behaviour. It may be seen, for example, as a high traffic volume in a short period of time, so it might be important to identify the parameters that allow the system to detect these examples. Some of these parameters could be the number of packages per time interval or the time interval between each package.

Regarding the ensemble classifier, as we can see in the results, given that two of the three classifiers always present very unfavorable results, the contribution of the *ensemble*, if any, is not significant.

The results are promising in some way, as the tools used in the attacks have produced patterns with some resemblance, thus making it possible for a behaviour-based IDS to use these algorithms to be able to identify a *zero-day* attack.

In scenarios 1 and 2, despite having a low True Positive Rate (TPR), the CLONALG algorithm stands out, together with the *ensemble*, as can be seen from the F1 values. The MLP algorithm has shown to be incapable of handling with this kind of traffic, only correctly identifying the overwhelming majority of normal traffic.

In contrast, in scenario 2, the LVQ algorithm presented the highest TPR in the scenario, despite failing to identify normal traffic (lowest True Negative Rate (TNR) value in the scenario).

In scenarios 3 and 4, we can depict the predominance of the MLP algorithm, with high F1 values, very close to 100% in scenario 3.

In scenario 3, as can be seen in the table 5, all algorithms correctly identified all malicious traffic (TPR = 1), which may be related to the similarity of traffic patterns generated by the respective tools. With regard to normal traffic, only MLP performs well, with TNR very close to 100%, while the other algorithms have a very residual identification.

In scenario 4, despite the great diversity of malicious traffic both in the training and testing phases, the traffic generated by the two tools in each type of attack has no significant advantage when compared to the results obtained in scenario 3. In fact, the performance of MLP, translated by the F1 value, drops by about 10%, whereas, in the other algorithms, there is little improvement.

6 Conclusions and Future Work

In this paper we have described a methodology to test bioinspired machine learning algorithms, against the recent and promising CSE-CIC IDS-2018 dataset. We described the dataset and the methodology used to process the four scenarios defined in each module. To fully automate the tests we have made available a tool developed with WEKA Java API.

We have sought to obtain statistical significance by running the tests ten times for each algorithm. The parameters set used in each algorithm was obtained empirically, combining the requirements of the algorithm itself and the data to be analysed.

In the first two scenarios, the highlighted algorithm is CLONALG, although the TPR is quite low, while MLP algorithm reveals poor performance. Despite correctly identifying the overwhelming majority of normal traffic, it clearly fails to identify malicious traffic. In the scenarios 3 and 4, the MLP performance is promising, with F1 and TPR values above 89%.

In addition to results obtained by each algorithm individually, an *ensemble* classifier was also implemented, which, using a majority voting strategy, had no significant influence in the final results. The future work includes the optimization of the parameters set and the processing of others datasets derived from CSE-CIC IDS-2018 dataset, with different attacks for training and testing.

References

[1] E. Frank, M. A. Hall, and I. H. Witten, "The weka workbench," in *Data Mining: Practical Machine Learning Tools and Techniques*, M. Kaufmann, Ed., 4th ed. 2016, ch. Online Appendix. [Online]. Available: https://www.cs.waikato.ac.nz/ml/weka/Witten_et_al_2016_appendix.pdf.

[2] L. N. de Castro and F. J. Von Zuben, "The clonal selection algorithm with engineering applications," in *Proceedings of GECCO*, editor, Ed., 2000, pp. 36–39.

[3] T. Kohonen, *Self-Organizing Maps*, ser. Springer Series in Information Sciences. Springer Science & Business Media, 2001, vol. 30. DOI: 10.1007/978-3-642-56927-2.

[4] F. Amato, N. Mazzocca, F. Moscato, and E. Vivenzio, "Multilayer perceptron: An intelligent model for classification and intrusion detection," in *2017 31st International Conference on Advanced Information Networking and Applications Workshops (WAINA)*, IEEE, 2017, pp. 686–691.

[5] A. Fuchsberger, "Intrusion detection systems and intrusion prevention systems," *Information Security Technical Report*, vol. 10, pp. 134–139, 3 2005. DOI: 10.16/j.istr.2005.08.001.

[6] M. Mahboubian and N. A. W. A. Hamid, "A machine learning based ais ids," *International Journal of Machine Learning and Computing*, vol. 3, no. 3, pp. 259–262, Jun. 2013. DOI: 10.7763/IJMLC.2013.V3.315.

[7] J. Kim, P. J. Bentley, U. Aickelin, J. Greensmith, G. Tedesco, and J. Twycross, "Immune system approaches to intrusion detection–a review," *Natural computing*, vol. 6, no. 4, pp. 413–466, 2007.

[8] J. Demšar, T. Curk, A. Erjavec, Č. Gorup, T. Hočevar, M. Milutinovič, M. Možina, M. Polajnar, M. Toplak, A. Starič, M. Štajdohar, L. Umek, L. Žagar, J. Žbontar, M. Žitnik, and B. Zupan, "Orange: Data mining toolbox in python," *Journal of Machine Learning Research*, vol. 14, pp. 2349–2353, 2013.

[2] https://github.com/paulo-ferreira-mcif/benchmarkids

Vessel Segmentation on Low-Resolution Retinal Imaging

Paulo Coelho[1,2]
paulo.coelho@ipleiria.pt
José Camara[3]
jcrcamara@hotmail.com
Hasan Zengin[4]
hasalp38@gmail.com
João M. F. Rodrigues[5]
jrodrig@ualg.pt
António Cunha[2,6]
acunha@utad.pt

[1] Escola Superior de Tecnologia e Gestão, Politécnico de Leiria
[2] INESC TEC - Instituto de Engenharia de Sistemas e Computadores, Tecnologia e Ciência Porto, Portugal
[3] Universidade Aberta, Porto, Portugal
[4] Mehmet Akif Ersoy University, Turkey
[5] LARSyS e ISE, Universidade do Algarve, Portugal
[6] Universidade de Trás-os-Montes e Alto Douro Vila Real, Portugal

Abstract

Retinal vessel segmentation process highlights a set of signals that will serve to aid for diagnosing of various retinal pathologies and lead to a more accurate diagnosis. This paper presents a framework for automatic vessel segmentation applied to lower-resolution retinal images taken with a smartphone equipped with D-EYE lens. A private dataset was assembled and annotated, and two CNN based models were trained for automatic localisation retinal areas and vessel segmentation. A Faster R-CNN that achieved a 96% correct detection of all regions with a Mean Absolute Error (MAE) of 39 pixels, and a U-Net that reached a Dice Coefficient (DICE) of 0.7547.

1 Introduction

Retinal imaging is a technique that allows recording digitally the rear of the eye. These are typically taken by expensive machines like fundus cameras, that produce high-quality and high-resolution retinal images for analysis. Then, with vessel segmentation, interference from other anatomical structures are filtered, helping to obtain the focus of interest on posterior segment structures of the eye. Manually segmenting retinal veins requires minutia, is a burdening task, time- and cost-consuming. Therefore, investigations of automatic or semi-automatic methods for vessel segmentation have been evolving to assist specialists [1]. However, the use of low-cost lenses, such as D-EYE [2], can bring several advantages such as greater portability, ease of use, greater patient comfort, lower costs and so can be an assessment for unprivileged or remote populations. The drawback is the lower quality of the photos obtained when compared to fundus cameras and as consequence not having the necessary sharpness when used in eyes with small pupils, in eyes with opacity of media (keratitis, cataract), or in very bright environments.

The latest trends in research show the extensive use of convolution neural networks (CNN) for the segmentation of retinal vessels and detection of the disease, beyond many other methods [1]. Nevertheless, these methods are all focused on segmenting vessels with high-resolution retinal images. There is still a lack of studies to evaluate the effectiveness of automatic methods to segment retinal vessels in this type of image. These low-resolution and low-quality retinal images create extra difficulties in the use of traditional vessel segmentation methods.

This paper presents a framework focused on the vessels segmenting on lower resolution retinal images taken with a smartphone equipped with D-EYE lens. The framework has two main steps: (A) The detection of the optic disc region using a Faster R-CNN and (B) Visible vessel segmentation made by U-Net, both trained with a customised dataset. The dataset was created with 26 retina videos around the optic disc, with lower-resolution images, and two annotated subsets, one with the localisation of the visible retinal area and other with vessel segmentation.

2 Methodology

This work is divided into two experiences (see the pipeline in Figure 1), applied to several datasets as follows.

Dataset

A dataset of 26 low-resolution videos of the optic papilla under myosis (undilated pupil) was captured from the left and right eyes of 19 volunteers. The videos were split into single images and organised in two different datasets: dataset1 (DS1), with a total of 6060 frames with 1920 x 1080 pixels to be used in the detection of the visible retinal area (from 18 videos were gathered 3881 frames to train, from 3 videos were gathered 776 frames to validation, and from 5 videos were used 1375 frames to test); and the dataset2 (DS2), with a total of 347 frames with 320x320 pixels pixels to be used in the segmentation of the retinal veins (from 2 videos were gathered 252 frames to train, from 1 video were gathered 40 frames to validation, and from 1 video were used 55 frames to test). Additionally, as dataset3 (DS3), a training set of retina public dataset [3] was used for pre-training a segmentation CNN. It is composed of 20 colour images with 565x584 pixels, resulting in 2810 patches with 80x80 pixels (for training, from 14 frames resulted in 1967 patches, for training from 3 frames resulted 421 patches and for testing, from 3 images resulted 422 patches).

Setup

In the first part of the experience, the detection of the retinal visible area (A) consists of computing the location of a rectangle defined by P1 and P2 (see Figure 1), that encloses the visible area in the image (the area of interest). In this case, input images have 1920x1080 pixels (from DS1), due to the D-EYE low lens aperture the area of interest has up to 320x320 pixels. For this purpose, it was used a Faster R-CNN model [4]. To evaluate the model, the Mean Absolute Error (MAE), which is a commonly used metric since it permits to measure the accuracy for continuous variables. In this particular case, four variables were used, two for the coordinates of the upper left corner (P1) and the other two for the lower right corner (P2).

In the second part of the experience (B), also depicted in Figure 1, the vessels segmentation was done within the detected retinal areas, with a U-Net [5] model pre-trained with the DS3 and tuned and evaluated with DS2. To measure the success of the model, it was used the Dice Coefficient (DICE), which is a relative metric that provides a similarity measure between predicted and ground truth segmentations.

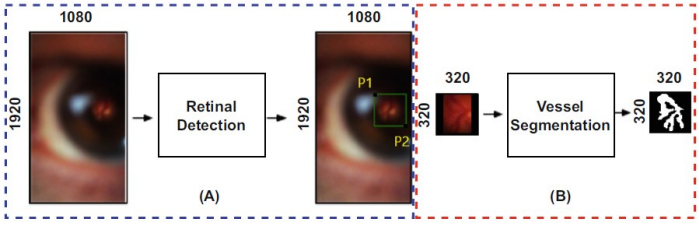


Figure 1: Pipeline diagram for the proposed low-resolution vessel segmentation framework.

For retinal detection (A), the Faster R-CNN model was used to detect retinal visible area detection. It was implemented in TensorFlow, with features pre-trained with Inception Resnet V2 and finetuned with the private dataset1. For augmenting dataset1, rotations were applied with 90-degree steps. The model was trained with default parametrisation: l2 regulariser of 0.01, truncated normal initialiser of 0.01, maxpool kernel size of 2, maxpool stride of 2, localisation loss weight of 2, objectness loss weight of 1, score converter Softmax, momentum optimiser with learning rate 0.0002, momentum optimiser value of 0.9.

In terms of retinal vessel segmentation (B), The U-Net model was implemented using Keras, with a TensorFlow backend. For training the U-Net model, it was used the binary cross-entropy as loss function, and Adam's optimiser with 10^{-3} learning rate based on Ange Tato and Roger Nkambou's work [6] used to achieve faster a stable convergence.

3 Results and discussion

The framework was evaluated for the retinal visible area detection and for vessels segmentation test sets.

The Faster R-CNN obtained results for retinal visible area detection are organised in 10 classification scores with intervals of 0.1 (represents the level of confidence of the detection), and can be seen in Table 1.

Table 1: Testset evaluation of the Faster R-CNN model for retinal visible area detection.

Classification score	Frequency	P1 MAE (pixels)*	P2 MAE (pixels)*	P1 and P2 MAE (pixels)*
0.0	30 (2%)	311 (414)	252 (321)	281 (371)
0.1	4 (0%)	46 (28)	41 (20)	43 (25)
0.2	12 (1%)	63 (41)	40 (31)	51 (38)
0.3	6 (0%)	47 (27)	30 (8)	38 (22)
0.4	7 (1%)	37 (23)	41 (26)	39 (24)
0.5	11 (1%)	76 (69)	50 (37)	63 (57)
0.6	11 (1%)	91 (60)	37 (20)	64 (52)
0.7	14 (1%)	61 (62)	35 (26)	48 (49)
0.8	29 (2%)	67 (50)	36 (20)	52 (41)
0.9	1,251 (91%)	47 (49)	28 (13)	37 (37)
Total images	1,375			

* MAE: mean (standard deviation)

The detection was very successful as 91% of the test images were detected with the classification score equal to or greater than 0.9 (see example in Figure 2, left). As the confidence score decreases, the MAE errors keep approximately constant until it reaches the score interval 0.0, where it increases for the mean of 281 and a standard deviation of 371 (see example in Figure 2, left). It was considered reasonable to use a threshold above 0.5 to accept the areas as valid-regions, achieving 96% of correct detected for all regions, with MAE of 39 pixels.

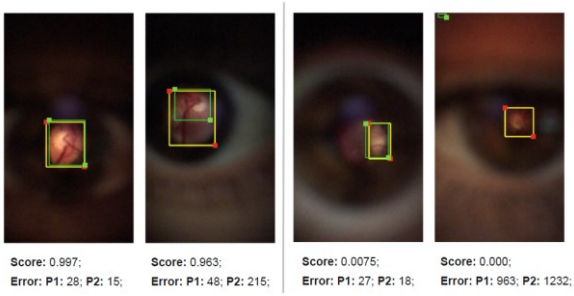


Figure 2: Example of the best and the worst sum of errors P1 and P2 in intervals 0.9 (left) and 0.0 (right). Yellow rectangles are the ground truth and the green are the model prediction.

The U-Net model was trained first with DS3 (Model 1), then trained at the junction of DS2 and DS3 testsets (Model 2) and later retrained Model 1 with DS2 (Model 3) to tune the network with D-EYE retinal data. The attained results are summarised in Table2.

Table 2: Results of the Model 1, Model 2 and Model 3

	Model 1	Model 2	Model 3
(DS3) testset	0.7824	–	0.5784
(DS2 & DS3) testset	0.7474	0.7312	–
(DS2) testset	0.4797	0.7547	0.5580

Model 1 has a reasonable Dice coefficient that seem adequate for the task (0.7824). Observing Figure 3 A), it can be seen that the best result (first column) has achieved a DICE of 0.935 in a patch where vessels are wide and well visible. The model predictions (row 3) have the same structure but seem wider than the ground truth (row 2). At the second column, it can be seen the worst prediction (DICE of 0.0512). At the original image, vessels are thin, almost imperceptible and quite different from the vessels expected to find in low-resolution images. It was selected another image patch with thin veins that seem to us more similar to the ones expected (third column). In this case, the predicted image preserves the structure; it also seem wider than the ground-truth and achieved a DICE of 0.8571. To observe how the Model 1 performs with the low-resolution images, it was evaluated in the DS2 testset, obtaining a low DICE value (0.4797). In the fourth and fifth columns, it can be seen the best and worst predictions. Both patch images are very dark, and veins are poorly visible - the image-patch of fifth is the poorest. The best-predicted segmentation (DICE of 0.8009) is very good, considering the visibility of the veins and though the difficulty of manually creating the ground-truth.

A) Model 1 B) Model 2

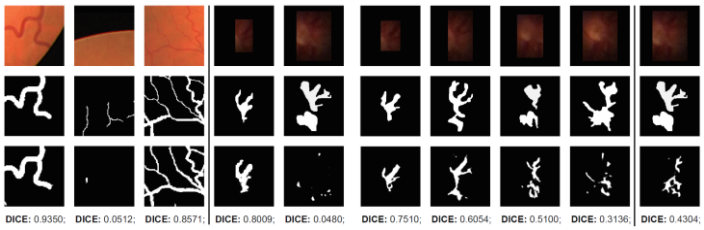


Figure 3: A) 1st row: original patches: the first three are from DS3 testset, respectively best, worst and reasonable predictions; and the last two from DS2 testset, respectively best and worst predictions. 2nd row, ground truth patches, and the 3rd row, Model 1 predictions. B) 1st row: original patches: the first four are from DS2 testset, respectively best, two in-between, and worst predictions; and for comparison, the worst case of Model 1 low-resolution prediction, is presented in the last column. 2nd row, ground truth patches. 3rd row, Model 2 predictions.

The Model 2 achieved a DICE of 0.7312 at the junction of DS2 and DS3 testsets that is lower than the obtained for Model 1 (DICE of 0.7474), but it achieved better on the DS2 testset (DICE of 0.7547). Examples of predicted images of Model 2 can be seen in Figure 3 B). To illustrate the Model 2 predictions of dataset DS2 testset, were chosen four images: the best prediction (DICE: 0.7510), two in-between predictions (DICE: 0.6054, DICE: 0.5100) and the worst prediction (DICE: 0.4304). For comparison with Model 1, column 5 has the prediction results of the worst-patch image predicted by Model 1 (see Figure 3 A), column 5). It can be seen that segmentations are much better: in the first two cases, the structure is all connected as in the ground truth, the other two (where veins are less visible in patch images) have several discontinuities in the structure. In column 5, one can see that Model 2 produces a much better segmentation (DICE: 0.4304) than produced by Model 1 (see Figure 3 A), lower right image).

The last tests made were for Model 3, by doing a posterior train of the Model 1 with DS2, but the results were worse than with Model 1.

4 Conclusions

In this paper, a framework for vessels segmenting on lower-resolution retinal images was proposed, evaluated, and the attained results were presented. A dataset of train models was assembled and annotated for automatic localisation of retinal areas and for vessel segmentation. For the framework, two CNN-based models were successfully trained, a Faster R-CNN that achieved a 96% correct detection of all regions with a MAE of 39 pixels, and a U-Net that achieved a DICE of 0.7547. This study is a precursor to future works to the determination of eye diseases, such as glaucoma and diabetes, applied to low-resolution images.

References

[1] Singh, N., Kaur, L.: A survey on blood vessel segmentation methods in retinal images. In: 2015 International Conference on Electronic Design, Computer Networks & Automated Verification (EDCAV), pp. 23–28. IEEE (January 2015).

[2] The Portable Ophthalmoscope for Your iPhone | Hand held fundus camera price| D-EYE for Humans | D-EYE, https://www.d-eyecare.com/en%20PT/product, Accessed: 2020-09-18

[3] Staal, J., Abramoff, M., Niemeijer, M., Viergever, M., van Ginneken, B.: Ridge-based vessel segmentation in color images of the retina. IEEE Trans. Med. Imaging 23(4), 501–509 (2004).

[4] Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans. Pattern Anal. Mach. Intell. 39(6), 1137–1149 (2017).

[5] Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015).

[6] Tato, A., Nkambou, R.: Workshop track -ICLR 2018 Improving Adam Optimizer, pp. 1–4 (2018)

Identifying Risky Dropout Student Profiles using Machine Learning Models

Sharmin Sultana Prite
sharmin.prite5@gmail.com
Teresa Gonçalves
tcg@uevora.pt
Luís Rato
lmr@uevora.pt

Departamento de Informática, Universidade de Évora, Portugal

Abstract

Student dropout prediction is essential to measure the success of an education institute system. This paper focuses on identifying the dropout risk at the University of Évora based on student's academic performance. Educational data was collected from four different programs, from the academic years of 2006/2007 until 2018/2019. After gathering the raw data, some data pre-processing was done aiming to build a dataset capable of being used by Machine Learning algorithms. Decision trees, Naïve Bayes, Support Vector Machines and Random Forests were evaluated, with the best model reaching an accuracy of around 96% when distinguishing between risky dropout and non-dropout students.

keywords: Machine Learning, Data Mining, Educational Data, Random Forest, Support Vactor Machines

1 Introduction

Nowadays, we live in the information era where acquiring data is easy and storing is inexpensive. Information is also the primary ingredient to generate new knowledge. The data mining can be applied in various real-life application like market analysis, education, and scientific exploration [6]. The use of data mining technique to analyze an educational database is absolutely expected to be a great benefit to the higher educational institutions.

Student dropout in Higher Education Institutions (HEIs) is, nowadays, a crucial concern for educators and managers. It also became a focus for researchers. Knowing, beforehand, the students at risk of dropping out allow higher education players to take measures that can contribute to an improvement in the institution success rate. Reasons for a dropout can be related to economical, social and psychological issues [1].

Anupama Kumar *et al.* [7] used a decision tree to help tutors identify the weak students and improve their performance before dropouts. Similarly, William C. Blanchfield [3] described a method of identifying college dropouts tested at Utica College of Syracuse University; he used multiple discriminant analysis to identify dropouts, reaching an accuracy of around 73%. Researchers from the University of Wuppertal developed an Early Detection System (EDS) [2] using administrative student data from a state and private universities to predict student success as a basis for targeted intervention; the EDS used regression analysis, neural networks, decision trees, and AdaBoost to identify student characteristics which distinguish potential dropouts from graduates. Yujing Chen *et al.* [4] developed and evaluated a survival analysis framework for the early identification of students at the risk of dropping out. In summary, existing approaches including logistic regression, decision trees and boosting showed good performance for early prediction of at-risk students and were also able to predict when a student will dropout. Given existing approaches, authors of this article tried different machine learning algorithms namely Decision trees (DT), Naïve Bayes (NB), Support Vector Machines (SVM) and Random Forests (RF) over academic data.This work uses student academic data from 4 different programs at the University of Évora to build classification models able to identify students at risk of dropping out.

The rest of the paper is organized as follows: Section 2 introduces the data used in this work, while Section 3 presents the developed work: data preprocessing, dataset generation, experimental setup, and results and their discussion. Finally, Section 4 concludes the paper and discusses future work.

2 Study Data

For this study, the students' full academic record was gathered. It considers four undergraduate study programs: Management, Biology, Computer Science and Nursing, during 13 academic years (from 2006/2007 to 2018/2019).

The student academic record includes information about course enrollments and corresponding results during the student university life: from the first year when student register at the university until graduation or dropout. Students were anonymized, and updates on study programs were considered. The list of information gathered from the information system are: *school year, degree, department, course code, course unit, regime, course credits, course name, edition, speciality, semester, time, type, student id, student type, mark, result, final status*.

3 Developed work

As previously mentioned, this work aims at creating a classification model using Machine Learning techniques to identify students at risk of dropping out so it could be used by authorities of HEIs to take possible actions aiming to reduce the number of dropouts. Figure 1 presents the block diagram of the developed work.

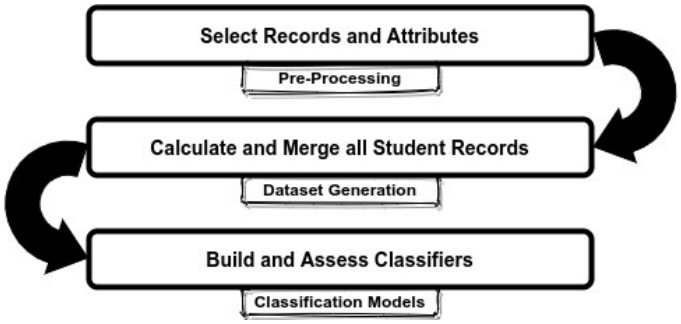


Figure 1: Developed work.

3.1 Pre-Processing

As already mentioned, student data was collected over a period of 13 years from four different undergraduate programs. In these programs, nursing is a four years program (totalling 240 credits), and the rest are three years of programs (totalling 180 credits).

The total number of enrollment records retrieved was 119407: 33731, 21328, 28689 and 35659 records for Management, Biology, Computer Science and Nursing, respectively. Information about the number of years taken to conclude the program is presented in Table 1.

Program	Min	Max	Avg	Stdev
Management	3	12	3.71	1.09
Biology	3	9	3.80	1.05
Computer Science	3	10	5.13	1.30
Nursing	3	13	4.25	0.64

Table 1: Information about the time taken to complete the programs.

Students enrolled at the university in the academic year of 2018/2019 were removed since at the time of data retrieval there was no academic record for them; this resulted in a total of 2934 students distributed as presented in Table 2.

From the data available at the University's information system, the following enrolment attributes were considered: *Academic_Year*, *Management*, *Biology*, *Computer Science*, *Nursing*, *Semester*, *Std_Id*, *Course*, *Credits*, *Mark*, *Final_Status*. Final_Status has two values: **S** means student pass the course, and **N** means student miss or fail the course. Course enrollment records without a value for Final_Status were removed because student enrolled the course but had not done any course activity.

Program	Number of students
Management	885
Biology	556
Computer Science	598
Nursing	895
Total	2934

Table 2: Number of students per study program

3.2 Dataset Construction

Using the retrieved data, and for each student, the annual student performance was calculated, and all the annual records were joint together to generate a single example; this example represents the academic path of a specific student.

The annual student performance is given by three attributes: the total number of enrolled and completed credits and average grade. This information was compiled for the student's five most recent academic years plus the performance calculated over the remaining student academic life. For students that successfully completed the program in less that 5 academic years, the values for attributes of oldest years were filled with zeros.

At the end, a dataset of 13 years composed by 21 attributes was built. Table 3 presents them.

Name	Number	Type
program_ects	1	int
program_name: man, bio, cs, nurse	4	bool (all)
year_0: enrol , avg_grade	2	int, float
year_1: enrol, complete, avg_grade	3	int, int, float
year_2: enrol, complete, avg_grade	3	int, int, float
year_3: enrol, complete, avg_grade	3	int, int, float
year_4: enrol, complete, avg_grade	3	int, int, float
year_rest: enrol, complete	2	int, int

Table 3: Dataset attributes.

A class label was then given to each example: success and unsuccess. The rule used was the following:

```
if registred = 2017 and completedCredit > 0
then SUCCESS
elseif registred < 2017 and completedCredit >= 210/150[a]
then SUCCESS
else UNSUCCESS
```

[a] 210 for nursing; 150 for other programs. This corresponds completing all except the credits of one semester.

The attributes and rules just described building the dataset were chosen considering a set of preliminary experiments that analysed other sets aiming to determine student success or unsuccess.

3.3 Classification Models

Four machine learning algorithms were used to build classifier models: Decision Tree (DT), Naïve Bayes (NB), Support Vector Machines (SVM), and Random Forest (RF). Weka 3.8.1 toolkit [5] was used for the experiments.

To tested the importance of the enrolled program and grade information, four different attribute subsets were used to build classification models:

- att_1: without *program_name*, without *avg_grade*
- att_2: with *program_name*, without *avg_grade*
- att_3: without *program_name*, with *avg_grade*
- att_4: with *program_name*, with *avg_grade*

The dataset was split into 70% of examples for training (2052 samples) and 30% for testing (882 samples). Then build the model using a training set and re-evaluated the model using the test set. To fine-tune the classifier algorithms, 10-folds cross-validation over the train set using the accuracy measure. Here, default parameter of all algorithms produce best results.

Table 4 shows the results obtained over the test set for each of the machine learning algorithms. As can be seen from the table, the overall performance by each algorithm over all the attributes is similar. The maximum difference of results is ranging from 0.67% to 1.71%, where RF has a minimum variation of 0.67%, and DT has a maximum of 1.71%. RF is outperforming all other algorithms by achieving 96.83% of accuracy.

Attributes	DT (%)	NB (%)	RF (%)	SVM (%)
Att_1	94.44	92.86	96.49	95.46
Att_2	94.90	92.74	96.15	96.15
Att_3	96.03	92.40	**96.83**	95.92
Att_4	**96.15**	**93.65**	96.60	**96.49**

Table 4: Accuracy results over test set.

Attributes	DT (%)	NB (%)	RF (%)	SVM (%)
Att_1	90.9	85.9	94.2	92.4
Att_2	91.7	88.4	93.7	93.6
Att_3	93.6	88.2	94.8	93.2
Att_4	93.8	89.9	94.4	94.2

Table 5: F-Measure Results over test set (Unsuccess class).

The F-measure results over unsuccess class of test set present in Table 5. The maximum difference of results is ranging from 1.1% to 4.0%, where RF has a minimum variation of 1.1%, and NB has a maximum of 4.0%. RF is out-performing all other algorithms by achieving 94.8% of F-measure.

From tables 4 and 5, it's not concluded that the best performance by RF is only achievable when all available attributes are not considered compared to the considering all attributes as the difference is the only 0.2% to 0.4%.

4 Conclusions and Future Work

This work presents an approach to identify dropout students by detecting risky profiles. It describes the available data, its preprocessing to generate a proper dataset and presents the results obtained using different machine learning algorithms. Using yearly enrollment information along with the study program and average grades an accuracy of around 96% for detecting risky dropout profiles was reached.

As future work, and to verify the results presented here we intend to enlarge the dataset to include more programs and, if possible, include student's personal, financial and social media information as attributes to improve the Machine Learning model.

Funding

This work was supported by the Erasmus Mundus LEADER (*Links in Europe and Asia for engineering, eDucation, Enterprise and Research Organization*) project.

References

[1] Jeff Allen, Steven B Robbins, Alex Casillas, and In-Sue Oh. Third-year college retention and transfer: Effects of academic performance, motivation, and social connectedness. *Research in Higher Education*, 49(7):647–664, 2008.

[2] Johannes Berens, Kerstin Schneider, Simon Görtz, Simon Oster, and Julian Burghoff. Early detection of students at risk–predicting student dropouts using administrative student data and machine learning methods. *CESifo Working Paper*, 2018.

[3] William C Blanchfield. College dropout identification: An economic analysis. *The Journal of Human Resources*, 7(4):540–544, 1972.

[4] Yujing Chen, Aditya Johri, and Huzefa Rangwala. Running out of stem: a comparative study across stem majors of college students at-risk of dropping out early. In *Proceedings of the 8th international conference on learning analytics and knowledge*, pages 270–279, 2018.

[5] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. The WEKA data mining software: an update. *SIGKDD Explorations*, 11(1):10–18, 2009.

[6] Jiawei Han, Jian Pei, and Micheline Kamber. *Data mining: concepts and techniques*. Elsevier, 2011.

[7] S Anupama Kumar and MN Vijayalakshmi. Implication of classification techniques in predicting student's recital. *Int. J. Data Mining Knowl. Manage. Process (IJDKP)*, 1(5):41–51, 2011.

Classifying Soil Type Using Radar Satellite Images

Sajib Ahmed[1]
jack6148@gmail.com
Teresa Gonçalves[1]
tcg@uevora.pt
Luís Rato[1,2]
lmr@uevora.pt
Pedro Salgueiro[1]
pds@uevora.pt
J. R. Marques da Silva[3,4]
jmsilva@uevora.pt
Filipe Vieira[4]
fsvieira@agroinsider.com
Luis Paixão[4]
lpaixao.agroinsider@gmail.com

[1] Departmento de Informática, Universidade de Évora, Portugal
[2] CIMA, Universidade de Évora, Portugal
[3] MED, Universidade de Évora, Portugal
[4] Agroinsider Lda., Évora, Portugal

Abstract

The growth of the crop is dependent on soil type, apart from atmospheric and geo-location characteristics. As of now, there is no direct and cost-free method to measure soil property or to classify soil type. In this work, we proposed a machine learning model to classify soil type using Sentinel-1 satellite radar images. Further, the developed classifier achieved 72.17% F1-score classifying sandy, free and clayish on a set of 65003 data points collected over one year (from Oct 2018 to Sep 2019) over 14 corn parcels near Ourique, Portugal.

Keywords: Remote Sensing, Soil Electrical Conductivity, Sentinel-1, Machine Learning, Random Forest

1 Introduction

Precision farming involves the collection of detailed information of mineral, nutrients, water, soil texture, cation exchange capacity, drainage conditions, organic matter level, salinity, and subsoil characteristics over farmland [3]. Over the last few decades, many new technologies have been developed for measuring soil properties, and one of such is using remote sensing techniques [2].

Sentinel-1 [7] is a synthetic aperture radar instrument (SAR) satellite that provides images in two different polarizations: VV (vertical transmit, vertical receive) and VH (vertical transmit, horizontal receive). It consists of a constellation of two satellites, Sentinel-1A and Sentinel-1B, which share the same orbital plane with a 12-day revisiting period.

In precision farming, detailed information about the spatial characteristics of farm operations like yield estimation, field attribute maps and forecasting harvesting date are made available to the farmer. This information is gathered using a wide array of electronic, mechanical and chemical sensors which leads to measure and map soil and plant properties. Soil Electro-Conductivity (EC) is one of the simplest, least expensive soil measurements available to precision farming today [8].

EC is the ability of a material to transmit (conduct) an electrical current and is usually expressed in miliSiemens/meter (mS/m). Soil EC is a measurement that characterizes soil properties which, in turn, affect the productivity of crops. These properties include water content, soil texture, soil organic matter (OM), depth to clay layer, the capacity of cation exchange (CEC), salinity, calcium and magnesium [4].

The objective of the present study is to build a classification model using machine learning algorithms that characterize soil types using Sentinel-1 radar images.

The rest of the paper is organized in the following sections: Section 2 introduces the data used in this work, while Section 3 describes the machine learning model, the experimental setup, experiments and results. Finally, Section 4 concludes the paper.

2 Data Set Construction and Characterization

The Electro-Conductivity value from a set of 14 parcels of corn fields (made available by Agroinsider [1]) was used as ground data points. These parcels are from Alentejo region with coordinates between (37°56'29.13" N , 8°22'21.95" W) and (37°55'32.44" N, 8°21'02.23" W). Figure 1 shows the Google View image of these 14 parcels. EC value was measured at 10-meter intervals resulting in a total of 65003 points.

Figure 1: Google view images of 14 parcels

Electro-conductivity real values were discretized, leading to three types of soil: sandy, free, and clayish. Table 1 presents the information about each type: the EC values interval and the number of points.

Soil Type	Value Range	Count
Sandy	$EC < 10mS/m$	24195
Free	$10mS/m \leq EC \leq 25mS/m$	31141
Clayish	$EC > 25mS/m$	9667

Table 1: Soil type information.

For each data point, along with the EC value, the respective latitude and longitude were also noted. With the collected coordinates, the corresponding values of VV and VH from the radar images were taken.

This radar data was collected from October 2018 to September 2019, the time span of one agricultural year. Since the Sentinel-1 revisiting time is 6 days, it resulted in a set of 60 pairs of values for each EC point measured. In this way, each soil point is characterized by 122 attributes: the soil type plus latitude, longitude and $60*2$ values of the radar images (60 dates and two polarizations: VV, VH). But latitude and longitude are not used as a parameter value in the ML algorithm.

Figure 2 represents the corresponding radar image for October 8, 2018 with VH polarization. And the variation of VH and VV value for one agriculture year (From Oct 2018 to Sep 2019) is shown in Figure 3.

3 Machine Learning Models

Three machine learning algorithms have been used to build classification models:

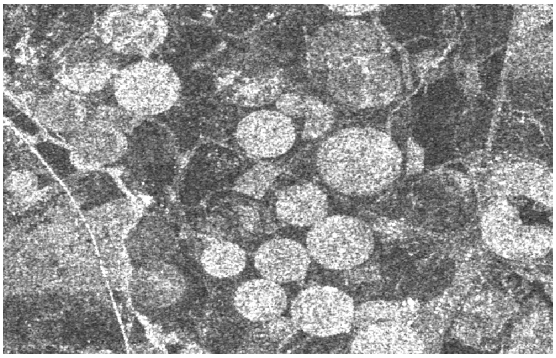


Figure 2: VH polarized radar image on 6^{th} October 2018

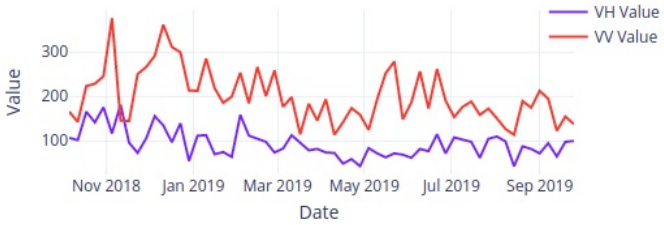


Figure 3: Variation of VH and VV values over a year in a specific point.

- Support Vector Machines (SVM) with a linear kernel.
- Random Forest (RF), a set of decision trees built from bootstrap samples of the training set where the candidate split in the learning process, is chosen from a random subset of the features.
- Extra Trees (ET), another ensemble classifier of decision trees, differs from RF in two points: each tree is trained using the whole learning sample and the top-down splitting in the tree learner is randomized.

3.1 Experimental Setup

A stratified train-test split was done over the dataset, with 80% for training (52002 samples) and 20% for testing (13001 samples).

We used Scikit-learn library [6] and RandomizedSearchCV [5] approach with 5-folds cross-validation to fine-tune the algorithms over micro-F1 measure. Parameters that produces the best results were: $nestimators = 189$, $max\_features = sqrt$, $max\_depth = 32$, $min\_samples\_split = 2$, $bootstrap = False$, $min\_samples\_leaf = 1$, and $criterion = gini$.

3.2 Experiments and Results

In order to evaluate the performance of the algorithms in this problem as well as the most relevant set of attributes, several experiments were carried out in a total of 153:

1. Algorithms: SVM, RF, ET
2. Time interval
   (a) 12 months
   (b) 3 months (Oct – Dec, Jan – Mar, Apr – Jun, Jul – Sep)
   (c) 1 month (Oct, Nov, Dec, Jan, Feb, Mar, Apr ..... Sep)
3. Polarization: VV, VH, VV + VH

These preliminary results made it possible to draw the following conclusions:

- Data set of 12 months time interval shows better results in performance measures: precision, recall and F1-Score.
- Compared to the other shorter intervals, performance increase between 2% to 3% in the F1-score measure, when compared to the results obtained with the April-June interval. The April-June interval presents the 2nd best F1-score values.
- The performance measure using only one of the polarization is similar. But some are gain (between 2% and 7% in the F1-score measure) when using both polarizations.
- Random Forest present the outperform than others based on the performance measures.

Table 2 details the results using Random Forest for the time span of 12 months. It presenting the best results in the three performance measures. So from 12 months time interval, several conclusions can be drawn from

Soil Type	Precision (%)	Recall (%)	F1-Score (%)
Sandy	79.70	70.15	74.62
Free	68.25	84.76	75.62
Clayish	80.17	41.21	54.44

Table 2: Performance of the Random Forest model over the test set.

the results:

1. it is possible to observe that the model behaves reasonably for sandy and free soils; precision is about 10% higher for sandy soils (almost 80%) but, on the other hand, free soils present 15% higher recall (about 85%);
2. concerning clayish soils, a high precision (about 80%) is obtained at the expense of a significantly low recall (about 41%); this difference affects F1-score, which fails to reach 55%, while for other types of soil the value is around 75%;

4 Conclusions and Future Work

This work presents a machine learning model to classify soil type using Sentinel-1 satellite images. The developed model, using Random Forests, is able to achieve 74.62%, 75.62% and 54.44% F1-score for sandy, free and clayish soils, respectively.

In future, to improve the results of this work, we will enlarge the dataset with more parcels having different crops, including more features from radar like the angle of incidence and timing for example.

Funding

This work was supported by NIIAA (Núcleo de Investigação em Inteligência Artificial em Agricultura) project, Alentejo 2020 program (reference ALT20-03-0247-FEDER-036981).

References

[1] Agroinsider an agricultural consulting company. `https://www.agroinsider.com/`, -. Accessed: 18 09 2020.

[2] Yufeng Ge, J Alex Thomasson, and Ruixiu Sui. Remote sensing of soil properties in precision agriculture: A review. *Frontiers of Earth Science*, 5(3):229–238, 2011.

[3] Robert Dwight Grisso, Marcus M Alley, David Lee Holshouser, and Wade Everett Thomason. Precision farming tools. soil electrical conductivity. -, 2005.

[4] Vr Ouhadi and Amir Reza Goudarzi. Factors impacting the electro conductivity variations of clayey soils. *Iranian Journal of Science and Technology Transaction B-Engineering*, 2007.

[5] David Paper and David Paper. Scikit-learn classifier tuning from complex training sets. *Hands-on Scikit-Learn for Machine Learning Applications: Data Science Fundamentals with Python*, pages 165–188, 2020.

[6] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[7] Paul Snoeij, Evert Attema, Malcolm Davidson, Berthyl Duesmann, Nicolas Floury, Guido Levrini, Björn Rommen, and Betlem Rosich. The sentinel-1 radar mission: Status and performance. In *2009 International Radar Conference" Surveillance for a Safer World"(RADAR 2009)*, pages 1–6. IEEE, 2009.

[8] Iulian-Florin Voicea, Mihai Matache, and Valentin Vladut. Researches regarding the electro-conductivity determination on different soil textures from romania, before sowing. *Bulletin of University of Agricultural Sciences and Veterinary Medicine Cluj-Napoca. Agriculture*, 66(1), 2009.

Prediction of pollution levels from atmospheric variables A study using clusterwise symbolic regression

Nikhil Suresh[1]
up201801861@fep.up.pt
Paula Brito[1]
mpbrito@fep.up.pt
Sónia Dias[2]
sdias@estg.ipvc.pt

[1] Faculdade de Economia
University of Porto & LIAAD INESC TEC,
Portugal

[2] Escola Superior de Tecnologia e Gestão
Instituto Politécnico de Viana do Castelo
Viana do Castelo, & LIAAD INESC TEC, Portugal

Abstract

This work performs statistical analysis of "Big data", considering the recent approach of Symbolic Data Analysis (SDA). The practical situation under study concerns the prediction of pollution levels in Senegal from atmospheric variables (meteorological indicators). The large number of records leads to the need of data aggregation. A temporal aggregation (by day) is made, where to each new unit (day) corresponds the interval of recorded values (minimum and maximum) in a given day. The symbolic data studied in this work is therefore interval data.

The objective was then to obtain symbolic regression models that allow explaining an objective interval-valued variable, the pollution level, as a function of explanatory interval-valued variables - the atmospheric variables. However, a single regression model is often not sufficient to adequately model the phenomenon under study. Thus, it was necessary to identify classes in the observed set and obtain a specific model appropriate for each class. To solve this problem, clusterwise regression for interval-valued data was developed.

1 Introduction

In classical data analysis, data is usually represented as an array where rows represent individuals and columns represent the variables (or attributes) describing them. It is possible to represent the data in a two dimensional array of n rows and p columns since a single value, numerical or categorical, is recorded for each variable and for each individual. However, when data is grouped to a higher level, the classical solution which is to use the mean, median or mode to represent each group leads to a loss of information, especially as concerns the variability present in each group. In such situations, SDA [1, 2] provides a framework to represent data with inherent variability, by using variables of special types. Among these representations, the focus in this work is on interval-valued data. A combination of existing dynamic clustering techniques and regression models for interval-valued data is proposed.

2 Problem: Predicting the levels of pollution in Senegal

The data under study consists of records of observations of atmospheric variables (meteorological indicators) and levels of pollution in Senegal, recorded from January 2006 to December 2010. The explanatory variables are wind speed, wind direction, air temperature and relative humidity, and the response variable is the particules concentration. The data was aggregated by day to form interval-valued variables recording the minimum and maximum values for each day. From the microdata, Table 1, the aggregation per day allows building an interval data array, as in Table 2.

Year	Month	Day	Hour	Min	Air Temp	Humitidy	...
2006	1	1	0	0	20.34	18.07	...
2006	1	1	0	5	20.30	18.09	...
2006	1	1	0	10	20.18	18.23	...
2006	1	1	0	15	20.14	18.30	...
⋮	⋮	⋮	⋮	⋮	⋮	⋮	...

Table 1: A snippet of Senegal meteorological indicators

The objective of this study is to predict the response variable, i.e., the particules' concentration, from the meteorological variables. However, a single regression model is often not sufficient to adequately model such relations. With the application of a clusterwise regression model for the interval data, we expect to obtain better results, by considering a partition of the time periods (days).

Year	Month	Day	Air Temp	Humitidy	...
2006	1	1	[20.13;20.34]	[18.07;18.30]	...
2006	1	2	[20.04;21.3]	[18.1;18.95]	...
⋮	⋮	⋮	⋮	⋮	...

Table 2: Senegal data snippet aggregation

3 The method

3.1 Interval Distribution (ID) regression model

Dias and Brito [3] proposed a new linear regression method for interval-valued variables known as the Interval Distribution (ID) regression model. In this approach, the intervals are represented by quantile functions taking into account the distribution within them. As it is usually the case in the literature, the Uniform distribution is assumed within each interval. Therefore, the quantile function that represents each interval is a linear non-decreasing function with domain $[0,1]$.

For each observation i of an interval-valued variable Y, $Y(i)$ is a interval $I_{Y(i)} = \left[\underline{I}_{Y(i)}, \overline{I}_{Y(i)}\right]$ where $\underline{I}_{Y(i)}, \overline{I}_{Y(i)}$ are the respective lower and upper bounds; $I_{Y(i)}$ may also be written as $I_{Y(i)} = \left[c_{Y(i)} - r_{Y(i)}, c_{Y(i)} + r_{Y(i)}\right]$, where now $c_{Y(i)}, r_{Y(i)}$ are the center and half range of the interval.

The quantile function that represents the interval $I_{Y(i)}$, when the Uniform distribution is assumed is written as $\Psi^{-1}_{Y(i)}(t) = \underline{I}_{Y(i)} + (\overline{I}_{Y(i)} - \underline{I}_{Y(i)})t$ or $\Psi^{-1}_{Y(i)}(t) = c_{Y(i)} + r_{Y(i)}(2t-1), \quad t \in [0,1]$.

Figure 3.1 represents the interval $I = [1,3]$ and the respective quantile function $\Psi^{-1}(t) = 1 + 3t, \quad t \in [0,1]$.

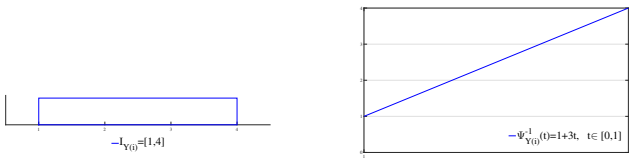


Figure 1: Graphical representation of the interval $[1,3]$ and respective quantile function.

The set of quantile functions defined from $[0,1]$ into R, with the usual operations of addition (between two quantile functions) and product of a quantile function by a real number, is not a subspace of the functions' vector space, but only a semi-vector space. The addition of two quantile functions raises no problem since the result is always a non-decreasing function. But the multiplication of a quantile function by a negative real number produces a function that is not non-decreasing, and hence cannot be a quantile function. Therefore, a problem arises when we multiply a quantile function representing an interval by -1, since we obtain a function that does not represent an interval.

As a result, when using quantile functions to represent intervals, the linear relation between interval-valued variables cannot be a direct adapta-

tion of the classical linear regression model. That is not possible because if the parameters of the model were negative, the quantile function predicted for the response variable Y could well turn out to be a decreasing function, i.e., not a quantile function. Applying non-negativity constraints on the model would guarantee a quantile function, but that would compel a direct linear relationship between the explanatory variables and the response variable, a too strict limitation. To allow for both direct and inverse linear relations between the response and the explanatory variables, Dias and Brito [3] proposed a method that considers not only the quantile function that represents the interval observation of each explanatory variable but also the quantile function that represents the respective symmetric interval. Therefore, the ID regression model allows predicting, for each unit i, the quantile function $\Psi^{-1}_{Y(i)}(t)$ from the linear combination of $\Psi^{-1}_{X_j(i)}(t)$ and $-\Psi^{-1}_{X_j(i)}(1-t)$, as follows:

$$\Psi^{-1}_{\hat{Y}(i)}(t) = a_0 + \sum_{j=1}^{p}(a_j - b_j)c_{X_j(i)} + \sum_{j=1}^{p}(a_j + b_j)r_{X_j(i)}(2t-1) \qquad (1)$$

with $t \in [0,1]$; $a_j, b_j \geq 0$, $j \in \{1,2,\ldots,p\}$ and $a_0 \in R$.

The non-negative parameters in the model are obtained by solving a quadratic optimization problem using the Mallows distance (see, e.g., [3]), used to measure the difference between the observed and the predicted quantile functions, for each unit $i, i \in \{1,\ldots,n\}$.

A measure Ω, similar to the classical coefficient of determination, was deduced for the ID regression model:

$$\Omega = \frac{\sum_{i=1}^{n} D^2_M\left(\widehat{Y}(i), \overline{Y}\right)}{\sum_{i=1}^{n} D^2_M\left(Y(i), \overline{Y}\right)} \qquad (2)$$

where $\bar{Y}$ is the symbolic mean of Y; $\widehat{Y}(i)$ and $Y(i)$ are the estimated and observed intervals of the interval-valued variable Y for each observation i. This measure, based on the Mallows distance D_M, measures the goodness of fit of the model, and ranges between 0 and 1.

3.2 Clusterwise Regression

The Clusterwise Regression model proposed in this work combines the dynamic clustering algorithm [4], with the ID regression model, considering a Uniform distribution within the intervals, in order to identify both a partition of the data units and the relevant regression models, one for each cluster. The steps of the algorithm to be followed are:

Step 1: Represent the interval data by quantile functions.

Step 2: Consider an initial partition of the given units.

Step 3: Fit a regression for each cluster using the ID Model.

Step 4: Re-assign each unit to the cluster that provides the best fit, as measured by the squared Mallows distance.

Steps 3 and 4 are repeated until convergence is attained and a local minimum of the sum of squares of the errors (measured by the Mallows distance) is obtained (or the fixed maximum number of iterations is reached).

The process may be applied varying the number of clusters K; for each fixed K, the implemented algorithm allows for different initial partitions, and selects the solution with lowest Total Error:

$$W = \sum_{k=1}^{K}\sum_{i \in C_k} D^2(Y(i), \widehat{Y}^k(i)) \qquad (3)$$

To select the best solution, across different K, we use the Weighted Coefficient of Determination [3],

$$\Omega = \sum_{k=1}^{K} \frac{n_k}{n}\Omega_k \quad \text{with} \quad \Omega_k = \frac{\sum_{i \in P_k} D^2_M\left(\widehat{Y}^k(i), \overline{Y}_k\right)}{\sum_{i \in P_k} D^2_M\left(Y(i), \overline{Y}_k\right)} \qquad (4)$$

where n_k is the number of observations in class k; $\bar{Y}_k$ is the (local) symbolic mean of Y in class k and $\widehat{Y}^k(i)$ is the estimated interval of $Y(i)$ obtained by the (local) regression model in class $k, k \in \{1,\ldots,K\}$.

Another measure used is the (adapted) Silhouette coefficient [5]:

$$S = \sum_{i=1}^{n} \frac{S(i)}{n} \qquad (5)$$

where, for each $i \in P_k$

$$S(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \qquad (6)$$

where the $a(i) = D^2_M(Y(i), \widehat{Y}^k(i))$ is the squared Mallows distance from unit i to its local estimate and $b(i) = \min_{\substack{l=\{1,\ldots,K\} \\ l \neq k}} D^2_M(Y(i), \widehat{Y}^l(i))$ is the minimum squared Mallows distance from unit i to the estimate provided by another class.

The final clusters may then be used to predict target intervals for new observations.

4 Results and Conclusions

The Clusterwise Regression method presented above was applied to the dataset described in Section 2 multiple times for different parameters. For each value of the number of clusters, 15 different initial partitions were analyzed. The algorithm was repeteadly applied until there was no increase in the value of the evaluation measure, or until the increase in the evaluation measure became negligible. Table 3 presents the best assessment measures received for each value of number of clusters. It was expected that the weighted coefficient of determination would rise with the number of clusters K. But it is no surprise that the rise would plateau after a certain value of K, in this case 5, for which the value of the weighted Ω attains 92%.

Nb. of clusters	Weighted Ω	Silhouete Coef.
2	0.7709	0.7963
3	0.8604	0.7187
4	0.9025	0.7052
5	0.9179	0.6892
6	0.9181	0.6840
7	0.9277	0.6692
8	0.9323	0.6438
9	0.9340	0.6717
10	0.9337	0.6679

Table 3: Model evaluation measures

The advantages of using a clusterwise regression model is that it fits one regression model for each cluster. Each cluster seems to have its own set of relevant regressors, with different values for these regressors. This provides a lot more flexibility than to fit a model for the entire dataset, which could dilute the effect of one specific regressor over a subset of data. In this case, with a global model we indeed obtain a worse fit, with $\Omega = 0.5685$.

References

[1] H.-H. Bock and E. Diday (Eds.). *Analysis of Symbolic Data. Exploratory methods for extracting statistical information from complex data*. Springer, Heidelberg, 2000.

[2] P. Brito. Symbolic data analysis: Another look at the interaction of data mining and statistics. *WIREs Data Mining and Knowledge Discovery*, 4(4):281–295, 2014.

[3] S. Dias and P. Brito. Off the beaten track: a new linear model for interval data. *European Journal of Operational Research*, pages 47–94, 2017.

[4] E. Diday and J.C. Simon. Clustering analysis. In *Digital pattern recognition*, pages 47–94. Springer, 1976.

[5] P. J. Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 1987.

Forecasting Ozone and Nitrogen Oxides for Air Quality Monitoring

César Bouças, cesarboucas@dei.uc.pt
Catarina Silva, catarina@dei.uc.pt
Alberto Cardoso, alberto@dei.uc.pt
Filipe Araujo, filipius@uc.pt
Joel Arrais , jpa@dei.uc.pt
Paulo Gil, pgil@dei.uc.pt
Bernardete Ribeiro, bribeiro@dei.uc.pt

Universidade de Coimbra
CISUC - Centro de Informática e Sistemas
FCTUC-DEI - Departamento de Engenharia Informática
Portugal

Abstract

Ozone (O3) and nitrogen oxides (NOx) emissions can harm ecosystems, agriculture and public health through their direct and indirect effects on the air quality. Thus, the ability to predict future concentrations of such gases is of strategic importance, especially in the current climate changing scenario. This work presents three methods to predict O3 and NOx concentrations for the upcomming 24 hours, given a sequence of past window of the same gas concentrations as input: a moving average, a linear regression and a Long short-term memory (LSTM) network that exibited the best result, being able to forecast NOx series with an average root mean squared error (RMSE) of 115ppb and mean absolute percentage error (MAPE) of 36% with respect to the ground truth series of the test set. The presented strategy was used to empower the NanoSen-AQM air quality platform.

1 Introduction

Gas concentrations observed at a regular interval of time (step) consist in a time series that can be used to predict future observations in a process called forecasting [1]. The forecast aim is to estimate how the observations will sequence into the future. Classical models used to forecast time series include ARIMA models, decomposition models and exponential smoothing [2]. Moreover, hybrid methods demonstrated advantages combining classical models with neural networks, such as in [3] that used exponential smoothing in conjunction with a Long short-term memory (LSTM) network and reached state-of-the-art results.

In this work, three methods are used to forecast hourly averaged NOx concentrations and two methods were used to forecast hourly averaged O3 concentrations. The number of future steps predicted was set to 24 and only the gas measurements were used as input to forecast future concentrations. Making the proposed methods simple enough to enable a smooth integration in the NanoSen-AQM online platform[1] [4].

2 Proposed approach

A moving average technique and a linear regression model were used as baseline, then a LSTM model was designed to enhance the performance. We avoided using extra features and specificities of the series in order to make our methods suitable to integrate and generalize well in the online platform dynamic environment.

2.1 Moving Average and Linear Regression

Two methods were used as baseline: a simple moving average since it produces predictions with no need for training, and a ordinary least squares regression, since it counts as a machine learning solution with low computational costs allied with resonable performance.

As the input for the Linear Regression, a sequence of 72 past measurements were used to predict the upcoming 24 measurements on Devito's data. For the Badajoz data, the lenght of the input sequence was reduced to 48 due the small number of examples.

The moving average were implemented as a simple arithmetic mean of past measurements. The mean counts as a future predicted step, thus, the method is repeated until the 24 future steps are predicted. Moving the window of past measurements towards the more recent values, one step at each iteration.

[1] https://nanosenaqm.dei.uc.pt/

We experimented several window lengths to calculate the mean, and 18 was the value that led to the best balance between error metrics and visual perception of the predictions.

2.2 Long short-term memory (LSTM) network

A NOx series forecast model was designed as a neural network whereas the input x_d sequence containing 72 past measurements is first transformed by a LSTM layer with 20 neurons (units) activated by a hyperbolic tangent function. Then by another LSTM layer with 8 rectified linear units and finally by an identity layer that outputs a sequence of 24 values that corresponds to the predicted future. Figure 1 illustrates the architecture.

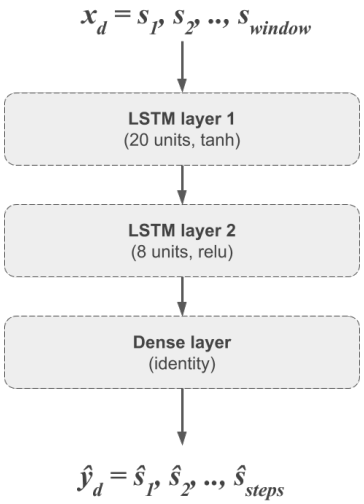


Figure 1: The neural network architecture with the number of neurons/units of each layer and its activation functions.

The network was trained through 50 epochs of backpropagation using gradient descent algorithm and mean squared error (MSE) as loss function.

The hyperparameters tuned at training were the number of hidden units and its activation functions. Whereas 20 and 8 hidden units with tanh and relu activation functions demonstrated to be sufficient to reach the best average results at training phase.

3 Experimental setup

3.1 Dataset

Series from two datasets were used to develop and test the proposed methods. As main source, averaged Nitrogen Oxides (NOx) concentrations recorded from March 2004 to February 2005 in Italy were used. This data is part of the Air Quality Data Set (Devito) [5] that is publicly available.

Ozone concentrations measured with reference sensors at Extremadura University campus (Badajoz) from September 21 to September 25 of 2017 were also used to deal with low data availability under the NanoSen-AQM data.

Method	RMSE (ppb)	MAPE (%)
LSTM	115.21	36
Linear Regression	131.34	41
Moving Average	215.86	84

Table 1: Results over Devito's NOx test set.

Method	RMSE (ppb)	MAPE (%)
Linear Regression	22.91	38
Moving Average	44.98	57

Table 2: Results over Badajoz's O3 test set.

3.2 Evaluation metrics

Given the the original ground truth series with the expected output ($y = s_1, s_2, ..., s_n$) and the series predicted by the model ($\hat{y} = \hat{s}_1, \hat{s}_2, ..., \hat{s}_n$). The RMSE measures the root average of the squares of the errors and its calculated as:

$$RMSE(y,\hat{y}) = \sqrt{MSE(y,\hat{y})} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(s_i - \hat{s}_i)^2} \quad (1)$$

The MAPE is calculated as a percentage:

$$MAPE(y,\hat{y}) = \frac{100}{n}\sum_{i=1}^{n}\frac{|s_i - \hat{s}_i|}{s_i} \quad (2)$$

3.3 Preprocessing

The gas concentration time series can be defined as a sequence of values v_i as such $D = (v_i)_{i=1..|D|}$. After removing empty rows and filling missing values with the last valid observation, the original values were rescaled since machine learning models tend to behave better when feature values are in a limited range near zero:

$$s_i = \frac{v_i - min(D)}{max(D) - min(D)} \quad (3)$$

In order to train a forecast model using supervised learning, we need to derivate from D, a new set D', consisting of pairs $(x_d, y_d)_{d=0..|D'|}$ where y_d is the expected output for an input x_d.

Having defined a constant *window* that is the number of steps used as input features. And a constant *steps* that is the number of future steps to predict:

$$(x_d, y_d) = ((s_i)_{i=1..window}, (s_j)_{j=window+i+1..window+i+1+steps}) \quad (4)$$

After derivating D' we splited it into train and test sets. The first was the major portion of the examples (75%) and was used to train the machine learning based methods. While the latter was left untouched and was used only to evaluate the models at the test phase.

4 Results

For each example in the test set, the trained models were used to make a prediction as well as the moving average was calculated. After doing this process over all the set, the evaluation metrics were calculated using the set of predicted values ($\hat{y}$) and the ground truth values (y) of the test set. Obtaining the final average errors for each test set: Badajoz and Devito.

Tables 1 and 2 sumarize the obtained metrics for the Devito and Badajoz test sets respectively. In the Badajoz case, the number of examples were insuficient to train the LSTM model.

Figure 2 illustrates an example from the Devito test set and the predicted outputs for this example. Offering a visual perception of the input, expected output and predictions of each method.

The results demonstrated that all the methods should be improved, expecially the moving average, which presented much worse metrics than the others methods despite the sufficient visual perception of it's predictions.

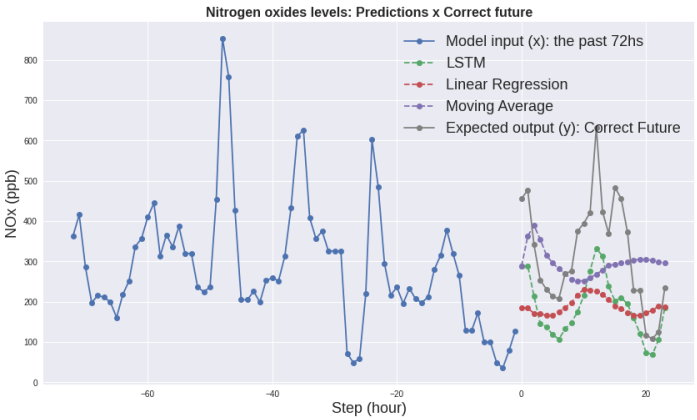


Figure 2: The predictions of the three methods for an example from the Devito's test set.

5 Conclusion and further work

The proposed methods were developed without using hand crafted features or series manipulation. Them demonstrated resonable performances, and were sucessfuly integrated into the NanoSenAQM online platform, where it is expected some generalization potential without requiring human intervention. Furthermore, the baselines showed to be attractive for it's simplicity and low memory consumption.

Results suggests that the seasonality of the series harm the performances, expecially of the moving average. Methods for automatic seasonality removal should be considered instead of classic manual removal methods, since the latter would not be suitable to be implemented as part of the online platform.

Besides removing the sasonality of the series, performance improvements can be reached by developing an exponential smoothing strategy within the LSTM such as [3]. The linear regression models might be benefited from exaustive hyperparameters search. Also, the moving average can be extended to a exponential implementation, giving greater importance to recent measurements in the inputs.

Finally, the use of informative features about temperature, humidity, wind and other factors that have impact in such gases behaviors could benefit the Linear Regression and the LSTM methods.

Acknowledgements

We acknowledge the Program Interreg-Sudoe of the European Union under grant agreement SOE2/P1/E0569 (NanoSen-AQM) and funding from the FCT, I.P., within CISUC Project UID/CEC/00326/2019 and CTS-UID/EEA/00066/2019.

References

[1] Rob J Hyndman and George Athanasopoulos. *Forecasting: principles and practice*. OTexts, 2018.

[2] Spyros Makridakis, Evangelos Spiliotis, and Vassilios Assimakopoulos. The m4 competition: 100,000 time series and 61 forecasting methods. *International Journal of Forecasting*, 36(1):54–74, 2020.

[3] Slawek Smyl. A hybrid method of exponential smoothing and recurrent neural networks for time series forecasting. *International Journal of Forecasting*, 36(1):75–85, 2020.

[4] Pedro Henrique Saraiva Lucas et al. *Development of the server for the NanoSen-AQM Project*. PhD thesis, Universidade de Coimbra, 2019.

[5] Saverio De Vito, Ettore Massera, Marco Piga, Luca Martinotto, and Girolamo Di Francia. On field calibration of an electronic nose for benzene estimation in an urban pollution monitoring scenario. *Sensors and Actuators B: Chemical*, 129(2):750–757, 2008.

Exploring a Siamese Neural Network Architecture for Drug Discovery

Luis Torres
uc2015241578@student.uc.pt
Department of Informatics Engineering, University of Coimbra

Bernardete Ribeiro
bribeiro@dei.uc.pt
Department of Informatics Engineering, University of Coimbra

Joel Arrais
jpa@dei.uc.pt
Department of Informatics Engineering, University of Coimbra

Abstract

Deep neural networks offer a great predictive power when inferring the pharmacological properties and biological activities of small molecules in drug discovery applications. However, in the traditional drug discovery process, where supervised data is scarce, the lead-optimization step is a low-data problem, making it difficult to find molecules with the desired therapeutic activity and obtain accurate predictions for candidate compounds. One major requirement to ensure the validity of the obtained neural network models is the need for a large number of training examples per class, which is not always feasible in drug discovery applications. This invalidates the use of instances whose classes were not considered in the training phase or in data where the number of classes is high and oscillates dynamically.

The main objective of the study is to optimize the discovery of novel compounds based on a reduced set of candidate drugs. We propose a Siamese neural network architecture for one-shot classification, based on Convolutional Neural Networks (CNNs), that learns from a similarity score between two input molecules according to a given similarity function.

Using a one-shot learning strategy, few instances per class are needed for training, and a small amount of data and computational resources are required to build an accurate model. The results achieved demonstrate that using a Siamese Deep Neural Network for one-shot classification leads to overall improved performance when compared to other state-of-the-art models. The proposed architecture provides an accurate and reliable prediction of novel compounds considering the lack of biological data available for drug discovery tasks.

Introduction

In drug discovery, we seek to maintain the desired properties of the main components of the molecules, preventing any structural deviation that might compromise their biological activity. Thus, the main objective is to discover novel compounds with optimal therapeutic effects, less toxicity, greater pharmacological activity, reduced risks for the organism, and better conditions of solubility and selectivity for the candidate molecules [1].

The feasibility of recognizing new compounds and their pharmacological analogs with a reduced set of biological data available for training remains an important challenge in compound prediction for drug discovery applications. Moreover, the identification of the class whenever a new group of molecules is observed, without requiring a periodic retraining and using only a few training examples per class, is crucial in drug discovery tasks.

Humans are able to learn multiple representations from a small number of examples, and then use the knowledge acquired to distinguishing new examples of these same representations, even if observed only once [2]. These idea of similarity gave rise to one-shot learning methods.

Instead of directly classifying a given instance, a one-shot learning model learns a similarity function that accepts two inputs, and returns a score that denotes the similarity between them. The learnt similarity rule allows to predict instances whose classes are unknown at training stage. The model learns a distance metric capable of distinguish two different inputs, and highlight the dissimilarities between them [3].

In the context of drug discovery, the application of a one-shot classification strategy improves the prediction of novel compounds whose classes are less-represented and only requires one example per class for training. Despite the size of the training set, a single molecule per class is needed for training. This molecule is used as a reference instance to compute the distance with any other molecule, while predicting a novel compound in one shot, according to the output similarity score generated between them. This similarity measure is the probability of both inputs belonging to the same class of molecular structures.

Model

A Siamese neural network built upon two parallel and identical convolutional neural networks is introduced as the proposed model approach. This network is compatible with a set of pairs of compounds provided for training. The model learns a similarity function and returns a distance metric applied to the output feature vectors from both siamese twins. This similarity measure allows the model to predict novel compounds in one shot, based on a reduced set of candidate molecules available for training.

Tox21 was the dataset used to extract SMILES (Simplified Molecular Input Line Entry System) for compound representation and encoding [4].

One-Shot Siamese Neural Network

We propose a Siamese Neural Network that accepts molecules organized in pairs. This model consists in two parallel and identical convolutional neural networks. Both Siamese twins are indistinguishable, since they are two copies of the same network and share the same set of parameters [5].

These parallel networks reduce their respective inputs to increasingly smaller tensors as we progress to the high-level layers. The difference between the output feature vectors is used as an input to the learnt similarity function. In a one-shot learning approach, one compound is established as a reference molecule and compared with different compounds expressing the probability of both belonging to the same class, according to a given similarity score *score*. The Siamese twins are symmetric neural networks, which means that the similarity score generated between d_1 and d_2 is equal to the score generated between d_2 and d_1. Thus, if we switch the order of the inputs of the Siamese network the returned output prediction would be the same:

$$score\,(d_1, d_2) = score\,(d_2, d_1) \tag{1}$$

This symmetry property is very important when learning a similarity metric. An architecture based on two parallel neural networks propagates two inputs through the same set of weights and the difference between the output feature vectors serves as an input to a similarity metric. This symmetry-based approach is less expensive and leads to a pairwise training which improves the model prediction accuracy.

Pairwise Training

A training set in which half are pairs of the same class and another half of different classes was considered. Since the Siamese neural network accepts pairs of molecules, the dataset size increases, given the number of possible combinations for the pairs of molecules available for training. However, we consider half of the pairs of the same class and half of the different classes for training. Therefore, the maximum number of possible combinations for compound pairs is the total number of possible pairs for compounds of the same class. If there are L examples each of Q classes, the total number of possible pairs of the same class is given by,

$$number\ of\ pairs = L \cdot \binom{Q}{2} = \frac{L \cdot Q!}{2! \cdot (Q-2)!} \tag{2}$$

The number of training instances increases in Q of a square factor and in L of a linear factor. The increase in the size of the training set reduces the effect of overfitting.

Model Architecture

The model architecture that maximizes performance is the one whose number of convolutional layers is 4, the number of filters in each layer is a multiple of 16, and in which the corresponding output features maps are applied to a ReLu activation function and to a maxpooling layer.

The output feature map of the last convolutional layer is flattened into a single vector that serves as an input to a fully connected layer with 1024 units. This layer learns a similarity function between two feature vectors by applying a distance metric to the learned feature map. It is followed by a dense layer that computes the absolute difference between the two output feature vectors. This value serves as input to a sigmoid function in the last layer.The predicted similarity score is given by,

$$score = sigmoid\ (\sum_i |\ v_{1,l}^i - v_{2,l}^i\ |) \tag{3}$$

v_1 and v_2 are the output feature vectors of the last convolutional layer of each Siamese twin, l the index representing the dense layer, i the index in each output feature vector and *sigmoid* the activation function. This defines a fully-connected layer for the network which joins the two Siamese twins and computes a distance metric over the feature space returning the similarity score between the two feature vectors.

The first Siamese twin returns the output feature vector for a given query molecule and the other returns an output feature vector for a molecule representing each one of the compound classes. This similarity measure is a probability, assuming a value between 0 and 1. If *score* is equal to 1, the probability of both compounds belonging to the same class is maximum. If *score* = 0, this probability is minimum according to the learnt similarity rule.

N-way One-Shot Learning Classification

The reduced amount of biological data available for training led us to adopt a new strategy to predict novel compounds using the proposed model. A one-shot classification strategy is applied to demonstrate the discriminative power of the learned features.

The Siamese network earlier described accepts pairs of compounds from a small training set D with a given number of N examples of encoded matrices of equal dimension and label l:

$$D = ((d_1, l_1),...,(d_N, l_N)) \tag{4}$$

The data for classification is organized in pairs, one example from a support set S and the other from a test set T. The support set consists of set of molecules representing each class selected at random whenever a one-shot learning task is performed. The support set has compounds representing each one of the compound categories and the test set has the test molecule of unknown class provided for classification.

In order to access the ability to make accurate predictions, a test instance d_j of unknown class is selected. Knowing that only one instance in our support set corresponds to that same class, the objective is to predict that class l belonging to D as the label l_i of an instance d_i.

Note that for every pair of input twins, our model generates a similarity score between 0 and 1 in one-shot. Therefore, to evaluate whether the model is really able to recognize similar molecules and distinguish dissimilar ones, we use an N-way one shot learning strategy (Figure 1). The test molecule is compared to N different ones and only one of those matches the original input. Thus, we get N different similarity scores $\{score_1,..., score_N\}$ denoting the similarity between the test molecules and those on the support set. This process is repeated across multiple trials, the model accuracy being determined as the percentage of correct predictions. Thus, in each trial, the pairs are organized for validation so that the first pair is a pair of instances of the same class, with the remaining pairs formed by compounds of different classes. If the pair of compounds of the same class (the first pair) gets the maximum similarity score, the model prediction is correct.

Over multiple trials, in each one-shot task, the Siamese network predicts which of the compounds present in the support set S most closely resembles the given test molecule in the test set T. The prediction *pred* corresponds to the pair (d_i,d_j) that returns the highest similarity score $score(d_i,d_j)$ in a one-shot trial with d_i the test molecule and d_j the support set molecule,

$$pred(d_i, S) = \arg\max(score(d_i, d_j)),\ d_j \in S \tag{5}$$

It is possible to verify that increasing N, more challenging it becomes to obtain a correct prediction and lower is the accuracy of the model. This is due to the fact that it is more difficult to obtain the maximum similarity score for the first pair due to the presence of a greater number of pairs in comparison at each trial.

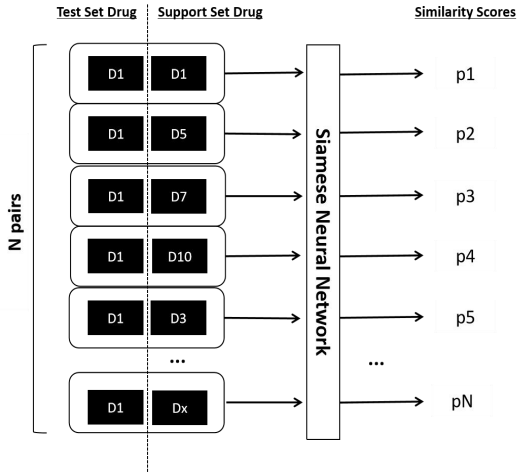


Figure 1: N-way one-shot classification using a Siamese neural network.

Results

In this paper, we propose a model able to predict novel compounds according to a degree of similarity between molecules. This metric is computed by a similarity function learnt by a one-shot Siamese neural network built upon two parallel and identical convolutional neural networks. To measure the model performance, the accuracy was determined using a N-way one-shot learning strategy described previously:

$$accuracy\ (\%) = \frac{number\ \ of\ \ correct\ \ prediction\ s}{number\ \ of\ \ trials\ \ per\ one-shot\ task} \tag{6}$$

The comparison of a given complex model with a set of simpler base models is a common strategy when assessing performance. Therefore, it was crucial to compare the proposed model with traditional machine learning approaches and simpler deep learning approaches (Table 1).

	N					
	2	3	4	5	7	10
Siamese Neural Network (validation)	94%	90%	84%	78%	70%	65%
Siamese Neural Network (training)	95%	92%	86%	84%	72%	70%
KNN	70%	55%	49%	43%	36%	30%
Naive Model	61%	43%	34%	31%	22%	19%
SVM	56%	42%	30%	24%	16%	12%
Random Forest	71%	58%	60%	44%	34%	20%
Multi-Layer Perceptron	76%	60%	36%	34%	22%	13%
Convolutional Neural Network	81%	70%	58%	46%	41%	39%

Table 1: N-way One-Shot Learning Accuracy Results.

References

[1] Altae-Tran, H., Ramsundar, B., Pappu, A. S., \& Pande, V. (2017). Low Data Drug Discovery with One-Shot Learning. ACS Central Science. https://doi.org/10.1021/acscentsci.6b00367.

[2] Lake, B. M., Salakhutdinov, R., Gross, J., \& Tenenbaum, J. B. (2011). One shot learning of simple visual concepts. In {Proceedings of the 33rd Annual Conference of the Cognitive Science Society}.

[3] Salakhutdinov, R. R., Tenenbaum, J. B., \& Torralba, A. (2012). One-Shot Learning with a Hierarchical Nonparametric Bayesian Model. JMLR Workshop and Conference Proceedings.

[4] A. Mayr, G. Klambauer, T. Unterthiner, and S. Hochreiter, "DeepTox: Toxic-ity prediction using deep learning,"Frontiers in Environmental Science, 2016.

[5] Koch, G. Siamese neural networks for one-shot image recognition. Ph.D. thesis, University of Toronto, 2015.

[6] N. R. C. Monteiro, B. Ribeiro, and J. Arrais, "Drug-Target Interaction Prediction: End-to-End Deep Learning Approach," IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2020.

Federated approaches for Remaining Useful Life prognosis

Raúl Llasag Rosero
rosero@dei.uc.pt
Catarina Silva
catarina@dei.uc.pt
Bernardete Ribeiro
bribeiro@dei.uc.pt

Universidade de Coimbra,
CISUC - Centro de Informática e Sistemas,
FCTUC-DEI - Departamento de Engenharia Informática,
Portugal

Abstract

Condition-Based Maintenance (CBM) using intelligent prognostic strategies that estimate Remaining Useful Life (RUL) has been applied in real scenarios to reduce maintenance costs and down times of machinery. When applied to aircraft maintenance, these models have also been developed in collaborative platforms that make use of data from similar components both in the same and different aircraft.

Even though RUL predictors have been presenting potential opportunities for developing federated scenarios after aggregating machine learning models, accuracy improvements of these models have not been evaluated yet, mainly due to absence of aircraft data.

In this work, we propose two collaborative federated approaches to determine RUL prognosis. The first approach aggregates models of equivalent subsystems located in the same airplane, while the second approach aggregates equivalent subsystems on different airplanes. We analyse two different systems from the aircraft: Integrated Cooling System (ICS) and Cabin Air Conditioning and Temperature Control System (CACTCS). We present the study of possible sensor data combinations according to the two proposed collaborative approaches.

1 Introduction

Prognostic is a very promising paradigm that permits strategies as Condition Based Maintenance (CBM) for reducing maintenance costs and downtimes of machinery such as aircraft systems [1]. CBM uses Prognostics and Health Management (PHM) techniques and metrics as Remaining Useful Life (RUL) to schedule maintenance tasks which are based in the duration left for a system before it fails [4].

RUL estimators are categorized into three types: model-based approaches, data-driven approaches, and fusion approaches [7]. Model-based approaches require physics machinery knowledge, while data-driven approaches analyze data using statistics and maths [1].

Strategies for developing data-driven approaches are based on the use of Artificial Intelligence techniques which have been deeply explored in the context of time series forecasting [4]. On the other hand, in order to reduce the distance between the estimation and the theoretical RUL, a recent collaborative paradigm named Federated Learning has been integrating machine learning models based on neural networks [6]. These evaluations have been done using virtual data engines as Turbofan but the aggregation of other equivalent airplanes subsystems has not been evaluated yet. Thus, in the present document, the aggregation of subsystems of the same airplane and different airplanes is considered.

2 State of the Art

The distance between RUL estimations and theoretical RUL for Turbofan engines has been reducing using direct computation approaches [6]. These approaches have been applied because degradation trends have been identified in specific time sensor data [5]. Nevertheless, due to the noise of sensors, finding a tren degradation requires of feature selection processes and the use of Health Indicator (HI) measures as the input for RUL computation [2, 3, 4].

The computation of virtual HI based on-flight phases aggregation [2] and feature selection under 3σ rule [4] has been presenting promising results for RUL computation of Boeing 787 Systems. However, the integration of HI estimators of equivalent subsystems of different airplanes has not to be done because federated scenarios on Boeing 787 datasets have not been identified yet.

3 Proposed Approach

Health estimator based in physics models have been limiting the scalability of PHM systems because collaborative scenarios have been depending on centralized processing architectures. So, estimators based in data-driven approaches through the use of neural networks have been gaining interest after collaboratively improving the accuracy of prognostic models under a newfangled parading named Federated Learning [6, 8]. Using the global loss function of the Equation 1, federated techniques based on Gradient Descent minimization improve the accuracy of equivalent prognostics systems in private aggregations of machine learning models [8].

$$F(w) \triangleq \frac{\sum_{j=1}^{N} n_j F_j(w)}{n} \tag{1}$$

Accuracy improvements in terms of distance between theoretical RUL and RUL estimations have done after iteratively averaging prognostic models (F_j) on a Federated Server, obtaining a central model ($F(w)$) which contains the knowledge of a defined number of federation participants (n) [6].

4 Experimental Setup

4.1 Datasets

For this work, data collected from the Cabin Air Conditioning and Temperature Control System (CACTCS) pack and the Integrated Cooling System (ICS) pack of Boeing 787 airplanes have been available as two datasets. The CACTCS pack is part of the Environmental Control System (ECS) and provides cabin temperature management and control, while the ICS pack is one of the three main packages which provide cooling flow to the primary electronics, the galley, and the forward cargo air conditioning system.

4.1.1 Cabin Air Conditioning and Temperature Control System

This dataset comprises data extracted from 13 different airplanes, in which data of sensors, faults and removals are useful for *a posteriori* RUL computation. Sensor data were extracted from 2 Packs systems using a sampling rate of 1Hz (sample per second). For each CACTCS pack, the data were retrieved from 45 different anonymized sensors. Here 23 sensors catch pack general information, while the other 22 sensors catch information of 2 equivalent Cabin Air Condition (CAC) components.

The faults data contains three types of failure reports that have been occurred during flights. Flight Deck Events (FDE) faults have been automatically generated during flights after presenting anomalies in sensors, while Aircraft Technical Log (ATL) and Predictive Maintenance (PM) faults have been identified by the maintenance team. Due to PM faults present removal/replacements dates given by technicians, we consider using these faults for HI prognosis.

4.1.2 Integrated Cooling System

This dataset comprises information collected of 17 Boeing 787 airplanes during 21 months. Similarly to CACTCS, ICS is composed of sensor data, failures and removals. However, FDEs are not reported over time, closing to use failure information provided by technicians.

ICS sensor data was retrieved from 70 anonymized sensors, which include 4 equivalent Supplemental Cooling Unit's (SCU), 2 SCU Motor Controllers (SCU-MC) and 2 PUMPs. However, as failures and removals are known only for SCUs, we propose to use only 9 sensors as input

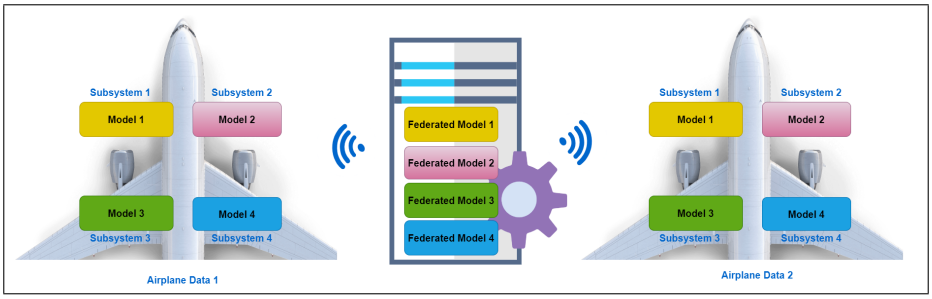


Figure 1: 1st Federated approach: Subsystems of different airplanes

of an SCU RUL predictor and reuse the four of them in collaborative approaches.

4.2 Dataset equivalent terminology

To aggregate HI estimators of equivalent components of CACTCS and ICS with federated techniques, Boeing 787 systems terminology of the Table 1 has to be generalized to describe the identified approaches.

Federated Learning	CACTCS	ICS
Subsystem	Component	Unit
Model	Comp. HI predictor	Unit HI predictor
Federated Model	Fed. HI predictor	Fed. HI predictor

Table 1: Generalized terminology used in collaborative approaches

CACTCS dataset is composed of sensor data obtained by different components, while ICS dataset contains units. Thus, in order to adopt the same terminology in a Federated context, systems correspond to the dataset and subsystems correspond to components or units, respectively.

5 Proposed Implementation

5.1 Federating subsystems of the same airplane

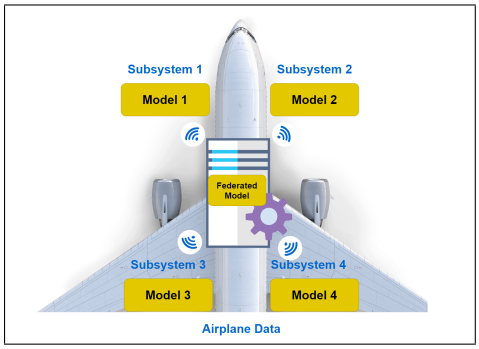


Figure 2: 2nd Federated approach: Subsystems of the same airplane

The approach illustrated in the Figure 2 assumes that some subsystems (same color boxes) per airplane are equivalent. In other words, the Remaining Useful Life of the 4 CAC components of CACTCS or 4 SCU units of ICS can be prognosticated using the same model.

Federation of equivalent subsystems does not require sharing models of different airplanes, but doing that, the prognosis accuracy of the federated model could be already improved.

5.2 Federating subsystems of different airplanes

In the Figure 1, the federation of equivalent subsystems (same color boxes) of different airplanes is illustrated. This approach generates one Federated Model per subsystem after aggregating models of different airplanes.

In the case of CACTCS, is assumed that the RUL of the 4 CACs of the left and the right packs can not be foreseen with the same model, i.e., the input sensors' data could be different for each subsystem. For ICS, this approach assumes that both SCU and PUMP are different.

6 Conclusions and Future Work

Due that the number of airplanes and failures are different for each dataset subsystem, the number of nodes for the both federated approaches are detailed in the Table 2. For the first federated approach, 10 CACTCS models and 9 ICS models contain the information of a same subsystem but located indifferent airplanes, while for the second federated approach, the number of federated constituents varies according each L-*th* and SCU-*th* subsystem. So, after developing and federating RUL predictors, improvements of federated approaches will be evaluated in future work.

Dataset	CACTCS	ICS
Subsystems	CAC (L1,L2,L3,L4)	SCU(1,2,3,4)
Airplanes	13	17
Failures	24	22
1st Fed. Approach	10	9
2nd Fed. Approach	6, 6, 4, 5	7, 4, 6, 2

Table 2: Federated cases for CACTCS ans ICS datasets

Acknowledgements

This work is part of a funded project from the European Union's Horizon 2020 research and innovation programme under grant agreement No769288.

References

[1] V. Atamuradov, K. Medjaher, P. Dersin, B. Lamoureux, and N. Zerhouni. Prognostics and health management for maintenance practitioners-review, implementation and tools evaluation. *International Journal of Prognostics and Health Management*, 8:31, 2017.

[2] D. Azevedo, A. Cardoso, and B. Ribeiro. Estimation of health indicators using advanced analytics for prediction of aircraft systems remaining useful lifetime. *Proceedings of the European Conference of the PHM Society*, 5(1):1–10, 2020.

[3] L. Guo, Y. Lei, N. Li, T. Yan, and N. Li. Machinery health indicator construction based on convolutional neural networks considering trend burr. *Neurocomputing*, 292:142–150, 2018.

[4] Y. Lei, N. Li, L. Guo, N. Li, T. Yan, and J. Lin. Machinery health prognostics: A systematic review from data acquisition to rul prediction. *Mechanical Systems and Signal Processing*, 104:799–834, 2018.

[5] E. Ramasso and A. Saxena. Benchmarking and analysis of prognostic methods for cmapss datasets. *Prognostics and Health Management Conference*, 2(5):1–15, 2014.

[6] R. LLasag Rosero, C. Silva, and B. Ribeiro. Remaining useful life estimation in aircraft components with federated learning. *Proceedings of the European Conference of the PHM Society*, 5(1):1–9, 2020.

[7] K. Tidriri, N. Chatti, S. Verron, and T. Tiplica. Bridging data-driven and model-based approaches for process fault diagnosis and health monitoring: A review of researches and future challenges. *Annual Reviews in Control*, 42, 2016.

[8] Q. Yang, L. Yang, T. Chen, and Y. Tong. Federated machine learning: Concept and applications, 2019.

Path planning by hybrid PSO-Splines algorithm

Paulo Salgado
psal@utad.pt

ECT- Departamento de Engenharias,
Universidade de Trás-os-Montes e Alto Douro, Vila Real, Portugal

Abstract

Trajectory planning is an NP-hard problem, and the computational effort of planning methods is usually high. In this paper is presented an algorithm for planning and optimizes trajectories. The trajectory is defined via Spline Cubic Polynomial interpolation, which represents continuous particles in a continuous search space. Therefore, to find the quasi-optimal trajectory, an adapted Particle Swarm Optimization algorithm is used, where particles are continuous spline functions. When applied to robots, this algorithm generates smooth motion trajectories, with two times continuously differentiable curves, and avoiding obstacles placed in the workspace. Thus, it can be used for autonomous robot path planning or transportation problems. It methods is also appropriate to contours image segmentation task. It was tested by hard 2D and 3D problems and the results demonstrated the effectiveness and performance of the algorithm.

1 Introduction

The scientific and industrial community have given much effort on creating optimization algorithms to solve global path planning problems. Also, there is a growing interest on path planning to be use by mobile robots, cars with autonomous driving, drones and industrial machines for many proposals [1-2]. This interest is now extended in contour task (segmentation) of digital image or to support analysis methods [3-4].

Path planning (PP) is only possible when the environment map is known. For instance, the robot's movement involves moving along a trajectory, starting from a specific point and finishing in an endpoint, passing through a sequence of points on its way and avoiding collision with other objects in the same workspace. Additionally, a robot's trajectory includes the motion of the robot with respect to time that is constrained by kinematic limits (e.g., joint velocity limit) and dynamic limits (e.g., torque constraint) of the robot joints and motors.

Hence, the goal is to plan and coordinate the motion of the robot axis. This is possible by fulfilling all constraints in order the robot passes all waypoints without any collisions and finally reaching its destination. The basic requirements of a good trajectory are the combination of smooth motion, a short and safety path. If these objectives are fulfilled, robots can efficiently improve its trajectory performance, namely by reducing the time of travel and the energy consumption [5].

Many trajectory-generation methods were proposed in the literature [6-7], which created motion trajectories for the robots considering several simultaneous criteria and constraints such as travelling distance, smoothness, security and feasibility [8], which it is almost a NP-hard problem. For the majority of the problems, it doesn´t exist an explicit solution or a deterministic method to solve them in an optimal and global way, taking into account all the constraints. To overcome these difficulties, the problem is usually simplified and therefore, the quality of the path solution is reduced. For many of these methods, a set of points with specific constraints is given and a path is generated from the combination of straight lines and circular arcs [9]. However, there is a curvature discontinuity at the straight line and at the circular arc joint. To cope with this problem, many researchers have modelled robot trajectories as piecewise quadratic or cubic Bézier curves [10]. Cubic polynomials splines have also been widely used as single curves to generate two times continuously differentiable curvature (trajectories), C^2, [11-13].

As I mentioned, in this paper I use the PSO algorithm to find a quasi-optimal trajectory through a feasible path in a complex environment, using a baseline smooth path based on cubic splines. This work is a step ahead of the algorithms proposed in [14], now with the improvement of the attraction/repulsive force of the Spline particles of the PSO algorithms. This method does not use traditional particle entities. Instead, they were replaced by continuous functions, namely by cubic splines, obeying certain C^2 continuity requirements throughout their domain and interpolating waypoints [15]. It is able to (iteratively) refine the path and thus find an efficient, collision free path in real time through an unstructured environment. This method is validated, and its performance evaluated through a set of simulations of hard and complex problems, with a great number of circular or spherical obstacles. The algorithm demonstrates a high success rate for all of the tested environments.

The rest of this paper is organized as it follows. Section 2 introduces the problem formulation for path planning. Section 3 introduces the PSO-Cubic Spline algorithm, PSO-CS, an optimization method in the space of continuous spline functions, and Section 4 presents experiments and result analysis. Finally, Section 7 summarizes the whole paper and presents the main conclusion.

2 Trajectory planning problem

Let a workspace W with n obstacles O_o, $o=1,\cdots,n$, and a set of trajectories $T_k=\left\{\left(t_j, X_j^{(k)}\right), j=0,\cdots,m+1\right\}$ where m are the number of waypoints with coordinates $X_j^{(k)} \in W$, the desired location of the robot at time t_j, specified in task space. It is assumed here that the positions of the starting and ending waypoints are provided.

For a given set of waypoints there is a unique piecewise-cubic trajectory, $S(t)$, that passes thr $y_j(t)=S_{y,j}(t)$ ough the points and satisfies a certain smoothness criterion. Specifically, let the Spline trajectory in $W \subset \boldsymbol{R}^3$ between times t_{j-1} and t_j as a cubic functions, one for each coordinate system: $x_j(t)=S_{xj}(t)$, and $z_j(t)=S_{z,j}(t)$, where S is a cubic polygon, i.e., $S_{r,j}=a_{rj}\Delta t_j^3+b_{rj}\Delta t_j^2+c_{rj}\Delta t_j+r_{j-1}\left(t_{j-1}\right)$ for $r\in\{x,y,z\}$. Founded a set of sequential waypoints, there exists a unique set of coefficients $\left\{\left(a_{rj},b_{rj},c_{rj}\right)_{r=\{x,y\}}\in\boldsymbol{R}^3\right\}_{j=1,\cdots,m}$, such that the resulting trajectory passes through the waypoints and has continuous C^2 profiles in the complete path. The waypoints are the visible solution of the optimization process performed by the PSO algorithm. However, in their evolutionary strategy, it takes into account not directly these points, but the complete path trajectory curve (spline) by evaluating its performance in the generated path. Truly, the PSO-CS, is an optimization method in the space of continuous spline functions.

3 Spline PSO algorithm

Particle Swarm Optimization, PSO, is inspired by the social behaviour of some biological organisms, especially the ability of some animal species to locate a desirable position in a given area. There are examples of this social behaviour on flock of birds and shoal of fish. This method is one of the optimization methods developed for finding a global optimal of some nonlinear function [18].

This method applies the approach of problem solving by a population of candidates solutions. Each solution consists on a set of parameters and represents a point in multidimensional space. The solution is called "particle" and the group of particles (population) is called "swarm". These particles move inside the search-space according to a few simple formula, they are guided by their own best known position in the search-space as well as the entire swarm's best known position, iteratively trying to improve a candidate solution with regard to a given measure of quality. So, each particle *i* is represented in a D-dimensional by the position vector $\vec{x}_i$ and it has a corresponding instantaneous velocity vector $\vec{v}_i$. It has a memory that tracks a best position of the previous iteration: the particle's optimal position *pbest* and the particle's global optimal position *gbest*. The particles are moving in *W*, under an action of one force that results from random combination of three effects: inertial, attraction of *pbest* and attraction of *gbest*. Under the effect of this force, the speed of particle is update in each iteration as:

$$\vec{v}_i(k+1)=\alpha\vec{v}_i(k)+c_1r_1\left(\vec{p}_{best,i}-\vec{x}_i(k)\right)+c_2r_2\left(\vec{p}_{gbest}-\vec{x}_i(k)\right) \quad (1)$$

where α is an inertia weight parameter, r´s are random numbers drawn from a uniform distribution in interval $[0,1]$, c_1 and c_2 are weights also designed as 'cognitive acceleration coefficient', respectively for local or global best position. The components values of $\vec{v}_i$ is restrict into the interval $[-v_{max}, v_{max}]$. Next, the position update rule (2) is applied:

$$\vec{x}_i(k+1) = \vec{x}_i(k) + \vec{v}_i(k). \quad (2)$$

In this work, the PSO method does not use traditional particle entities. Instead, they were replaced by continuous functions, namely by cubic splines. Cubic Spline that obeys a certain C^2 continuity requirements throughout their domain and interpolating waypoints defines the path. The arcs of Spline are subject to a force of repulsion by obstacle objects. This fourth component of the force that moves the particle in *W*.

Despite the differences, for simplicity of analysis, we are going to use the same designations indistinctly. S_i represents the i^{th} Spline particle, with waypoints X_i in *W*, that it have an instantaneous velocity vector function $\vec{u}_i(t)$. $\vec{U}_{ij} = \vec{u}_i(t_j)$ is the velocity vector of the j^{th} waypoint of spline particle i with travel trajectory time t_j. Under the influence of near objects, the arcs of Spline are subjects to a repulsion force. This force has a higher magnitude value if the arc spline is inside object body and, consequently, a less magnitude value for spline arcs further from the object. It has a null value for distances greater than the distance of safety margin. This tensor or strain force that propagates across Spline curve to their waypoints results in the force R_i. This joins to the other PSO force, which represents another right term of equation (1).

4 Tests and results

I tested the PSO-CS algorithm for robot path planning in ®MATLAB platform in two test-examples, the 1st in two-dimensional space, 2D, and the 2nd in three-dimensional space, 3D. The workspace has a shape like a square/cube with edge length of 100 units. Inside there are 30 circular/spherical obstacles with radius length of 10 units. In the first example, the circles are random placed on workspace while in the 2nd text-example a wall, made by 24 spheres and a hole at its centre that split the workspace in two zones. Three more spheres are randomly placed in each one-sided zones. We assume that the start position of the robot is in origin, $X_0 = [0,0] / X_0 = [0,0,0]$, and the end position is in opposite vertices, i.e $X_{m+1} = [100,100] / X_{m+1} = [100,100,100]$.

PSO-CS uses Spline curves as particles. It adjusts the randomly placed waypoints of Splines population, where the particles are the coordinates of these points, in a discrete optimization process. The spline particles have 5 waypoints. One hundred 'particles' have been used for simulations with 100 iterations. For both environments, the PSO-CS was tested. The best Spline path is recorded at each iteration. At the end of the process, the best Spline trajectory is shown, as well as the final population and the best performance evaluation in each iteration. Figure 1 and Figure 2 show the results, respectively, for the 2D and 3 D test-examples. The red line shown in figures represents the optimum path generated by the algorithms and the filled circles/spheres represent obstacles.

From simulations results, we can conclude that PSO-CS algorithm efficiently finds a collision-free path between the initial and destination points, the global solution with a quasi-optimal performance value.

5 Conclusions

This paper presents an improved version of the PSO-CS algorithm for global path planning. In order to get smoother planned paths, it uses a Cubic-spline smoothing technique. It was tested for robot motion planning problem, which it is treated as an optimization problem. Random obstacles are place on the 2D and 3D workspaces. The PSO-CS, in each iteration, try's to find a feasible Spline curve with the best performance, defined by appropriated waypoints. The result is a smoother planned path, which it is a quasi-optimal trajectory. Experimental results show that it is an excellent method for path planning, generating collision free and smooth trajectories with shorter paths length.

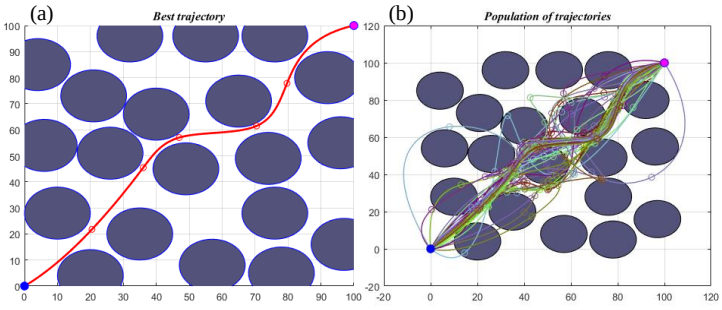


Figure 1: PSO-CS: (a) Best trajectory; (b) Population of splines.

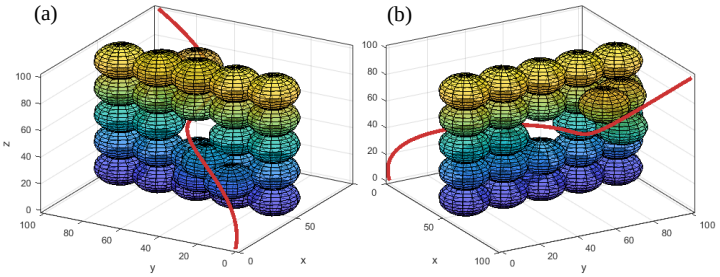


Figure 2: PSO-CS trajectory. View: (a) El. 30º, Ez. -60º and (b) Ez. 60º.

References

[1] Luigi Biagiotti and C. Melchiorri, "Trajectory Planning for Automatic Machines and Robots", Springer; 2008.

[2] Steven M. LaValle, Planning Algorithms, Cambridge Univ., 2006.

[3] R. Delgado-Gonzalo, M. Unser, Spline-based framework for interactive segmentation in biomedical imaging, IRBM, Vol. 34 (3), pp. 235-243, 2013. https://doi.org/10.1016/j.irbm.2013.04.002.

[4] L. Stanberry, J. Besag, Boundary reconstruction in binary images using splines, Pattern Recognition, Vol. 47(2), pp. 634-642, 2014.

[5] Steven M. LaValle, Planning Algorithms, Cambridge Univ., 2006.

[6] Serdar Kucuk, Optimal trajectory generation algorithm for serial and parallel manipulators, Robotics and Computer-Integrated Manufacturing, Vol. 48, pp. 219-232, 2017.

[7] B.K. Patle and at. al., A review: On path planning strategies for navigation of mobile robot, Defence Tech., Vol. 15(4), pp. 582-606, 2019.

[8] Prasenjit Chanak, and at.al., "Obstacle Avoidance Routing Scheme Through Optimal Sink Movement for Home Monitoring and Mobile Robotic Consumer Devices", IEEE Trans. on Consumer Electronics, Vol. 60, No. 4, pp. 596-604, Nov. 2014.

[9] Yi Fang and at. al., Smooth and time-optimal S-curve trajectory planning for automated robots and machines, Mechanism and Machine Theory, Vol. 137, Pages 127-153, 2019.

[10] Christian Scheiderer and at. al., Bézier Curve Based Continuous and Smooth Motion Planning for Self-Learning Industrial Robots, Procedia Manufacturing, Vol. 38, pp. 423-430, 2019.

[11] T. Berglund and at. al., "Planning smooth and obstacle-avoiding B-spline paths for autonomous mining vehicles," IEEE Trans. Autom. Sci. Eng., vol. 7, no. 1, pp. 167–172, Jan. 2010.

[12] M. Elbanhawi and at. al., Continuous path smoothing for car-like robots using b-spline curves, J. Int. Rob. Syst., vol. 80(1), pp. 23–56, 2015.

[13] R. Cowlagi and P. Tsiotras, "Curvature-bounded traversability analysis in motion planning for mobile robots," IEEE Trans. Robot., vol. 30, no. 4, pp. 1011–1019, Aug. 2014.

[14] P. Salgado, and at. al, "Hybrid PSO-cubic spline for autonomous robots optimal trajectory planning," 2017 IEEE 21st Int. Conf. on Intelligent Engineering Systems (INES), Larnaca, pp. 131-136, 2017, doi: 10.1109/INES.2017.8118542.

[15] G. Liu and S. Wu, "A Novel Optimized Trajectory Planning Method via Spline Function Theorems," 2018 IEEE 4th Inf. Tech. and Mech. Eng. Conf. (ITOEC), China, pp. 745-749, 2018.

Federated Learning Optimization

Miguel Fernandes
mfernandes@student.dei.uc.pt
Joel P. Arrais
jpa@dei.uc.pt
Catarina Silva
catarina@dei.uc.pt
Alberto Cardoso
alberto@dei.uc.pt
Bernardete Ribeiro
bribeiro@dei.uc.pt

University of Coimbra
CISUC - Centro de Informática e Sistemas
FCTUC-DEI - Departamento de Engenharia Informática
Coimbra, Portugal

Abstract

In a recent approach defined as Federated Learning (FL), a single model is shared between a server and the clients instead of the data itself, reducing the amount of data transferred. In addition, FL attenuates the privacy concerns since each model is computed locally by their respective client and only the model is shared.

Federated Learning is still a recent technology and, as such, much research is yet to be done. This work presents the proposal and implementation of two Federated Learning algorithms and comparison with state of the art.

1 Introduction

Federated Learning (FL) [1] is a rising decentralized learning technology. While conventional Machine Learning methods require data to be centralized, Federated Learning allows multiple clients to learn a shared global model without needing to send their local data to a server. In addition, all of the model's training is done locally and coordinated by a central server.

In a FL setting, the clients receive a shared model from the server θ_t and train it with data which is only accessible to it. Afterwards, each client sends the updated models to the server. In the server, the uploaded models are aggregated in order to form a new model.

This work proposes two new methods which outperform the state of the art (Federated Averaging and Federated Proximal) of Federated Learning: Federated Directional Congruent Learning (FedCong) and Federated Momentum (FedMom). While the first is based on the directions of the models' updates, the second algorithm is based on the momentum of the global model.

1.1 Federated Averaging

Federated Averaging (FedAvg) [1] is a FL algorithm which generates the global model by periodically averaging the clients' locally trained models [1].

The algorithm starts by initializing a global model θ_t. Afterwards, at the t Communication Round (CR), the server selects a random subset of clients, K, and uploads the current global model to the clients. The chosen clients then train θ_t by performing stochastic gradient descent (SGD) locally for E epochs. Lastly, the clients upload the resulting model to the server where they are aggregated using a weighted average given by:

$$\theta_{t+1} = \sum_{k \in K} \frac{n_k}{n} \theta^k_{t+1} \tag{1}$$

where n is the sum of all clients' local data n_k. It is empirically shown in the work by H.Brendan McMahan [1] that the tuning of the number of local updates is of major importance for FedAvg to converge. It is clear that more local updates cause the model to be fitter for the local optimization problem and move further away from the initial model, possibly causing divergence.

1.2 Federated Proximal

Federated Proximal (FedProx) [4] was developed with the purpose of restricting the amount of divergence of the local model with regard to the global model, removing the need for heuristically limiting the number of local updates.

FedProx is similar to FedAvg with the difference being that each client local optimizer minimizes the objective given by:

$$\min h_k \quad \text{where} \quad h_k = F_k + \frac{\mu}{2} ||\theta_t - \theta^k_{t+1}||^2 \tag{2}$$

where $\frac{\mu}{2}||\theta - \theta^t||^2$ corresponds to the *proximal term*, which reduces the effect of local updates by making the local model θ^k_{t+1} closer to the global model θ_t. A cautious reader will note that if $\mu = 0$, then this algorithm is the same as the FedAvg algorithm.

2 Proposed Algorithms

The next sections contain the new Federated Learning algorithms proposed and implemented in this work, namely Federated Congruent Directional Learning (FedCong) and Federated Momentum Learning (FedMom).

2.1 Federated Congruent Directional Learning

In this section, the FedCong algorithm will be presented. This algorithm tries to mitigate a problem in FedAvg. As it was previously explained, in FedAvg, the bigger the number of updates, the more fitted the model is to the local optimization problem, potentially causing divergence. This divergence can lead to a decay in the model's convergence speed.

The FedCong algorithm was developed taking these facts into consideration. It is similar to the other methods with the main difference being that at the server, for each local model θ^k_{t+1} received, each weight w^k_{t+1} update for the local problem is analysed. On the one hand, in case $w^k_{t+1} < w_t$, then it can be concluded that w^k_{t+1} had a negative update. On the other hand, in case $w^k_{t+1} > w_t$, then it can be concluded that w^k_{t+1} had a positive update.

After this process in completed, for each weight this algorithm calculates the number of positive and negative updates of all the local models. Afterwards, one of three possible situations will occur:

$$\begin{cases} \text{if} \quad P \geq K * \alpha, & \text{then average the positive clients} \\ \text{if} \quad N \geq K * \alpha, & \text{then average the negative clients} \\ \text{otherwise}, & \text{then average all the clients} \end{cases} \tag{3}$$

where K is the number of selected clients, P and N are the number of positive and negative updates for a specific weight, respectively, and $\alpha(0,1)$ is a control parameter which specifies the minimum number of positive or negative updates that are necessary to average a weight using only the positive or negative updates.

2.2 Federated Momentum Learning

The FedMom algorithm was inspired by the Momentum optimizer [3, 5], having the objective of maximizing the training speed of the FedAvg algorithm. Momentum is known to help the gradient vector pointing to the right direction, damping oscillations and taking more straightforward paths to the local minimum.

The following equations (4, 5, 6) show how the momentum update can be reformulated into a Federated Learning setting. Firstly, FedMom

starts by initializing a global model, θ_t. Afterwards, similarly to FedAvg, the server selects K clients and uploads the current global model, θ_t. Subsequently, the clients selected take one or multiple SGD steps locally as follows:

$$\theta_{t+1}^k = \theta_t - \eta a_k \quad where \quad a_k = (g_e^k + g_{e-1}^k + ... + g_0^k) \tag{4}$$

where θ_t^k is the local model of the $k-th$ client, η is the learning rate, t represents the global model's timestamp, e represents the local model's timestamp, g_e^k are the error gradients of θ_e^k, and a_k is the sum of all the gradient updates of the $k-th$ client's local model.

Afterwards, the clients upload the resulting model θ_{t+1}^k to the server. In the server, for every θ_{t+1}^k the server calculates the local update. Then, it calculates the global update of the model using each local update as follows:

$$\begin{aligned} -\eta a_k &= \theta_{t+1}^k - \theta_t \\ \alpha &= \sum_{k=1}^{K} \frac{n_k}{n}(-\eta a_k) \end{aligned} \tag{5}$$

where α is the global model update, n represents the total number of data points and n_k represents the number of data points of the $k-th$ client.

After this process is completed, the server stores the momentum variable and updates the global model as follows:

$$\begin{aligned} v_{t+1} &= \delta v_t + \alpha \\ \theta_{t+1} &= \theta_t + v_{t+1} \end{aligned} \tag{6}$$

where δ is the momentum term. As it can be observed, the momentum update (v_{t+1}) is calculated by summing a fraction of the previous update and the global model's update. Afterwards, the global model is updated by summing the previous model with the momentum update.

3 Experimentation

In this section, the experimental results of the Federated Learning optimizers will be presented. The algorithms tested were: FedAvg, FedProx, FedCong and FedMom.

In the scope of this work, the baseline model considered is the FedAvg since it is the most widely used algorithm for Federated Learning. In addition, FedProx is also evaluated. The primary objective of this work is that FedCong and FedMom outperform FedAvg by increasing the convergence speed of the models while maintaining the Mean Absolute Error (MAE). The best algorithm is the one whose CR of stabilization is the lowest without increasing the MAE.

The experimentation was done using the Turbofan Dataset. The Turbofan dataset is composed of four sub-datasets (FD001, FD002, FD003 and FD004). Each dataset has a time series readings of 26 features, such as Operational Settings, unit number, time indicator and 21 sensors' values regarding the turbofan engine components. At the start of each series, the system operates in a healthy condition until some point in time where it enters a failure state and can no longer function. This degradation is captured by the time indicator feature.

In addition, after some literature review, it was concluded that only 14 out of the 21 sensors presented valuable information to be used as input features.

For each dataset, J clients were created so that each client has exactly two time series. As such, each client only has a small portion of the data. In the following experimentation, for each CR, K clients were randomly selected from the J clients. The value of K for these experiments was set to 20.

Each client of the FL setting is represented by a Feed Forward Neural Network which has the objective of predicting the system's health percentage. This neural network takes as input the preprocessed features' values. The network architecture is as follows: three hidden layers with 20, 30 and 20 neurons, respectively, with *tanh* activation functions; the output layer is a single neuron with a *sigmoid* activation function, which represents the predicted system's health percentage. The local optimizer used was SGD and the error function used was the Mean Square Error (MSE).

In Figure 1, the results of the proposed methods are presented for the four different datasets (FD001, FD002, FD003 and FD004). The step size η is similar between all the algorithms. The α and δ from FedCong and FedMom, respectively, were tuned in order to obtain the best performance. It can be observed that the two new proposed algorithms converge faster than the others while maintaining the same MAE.

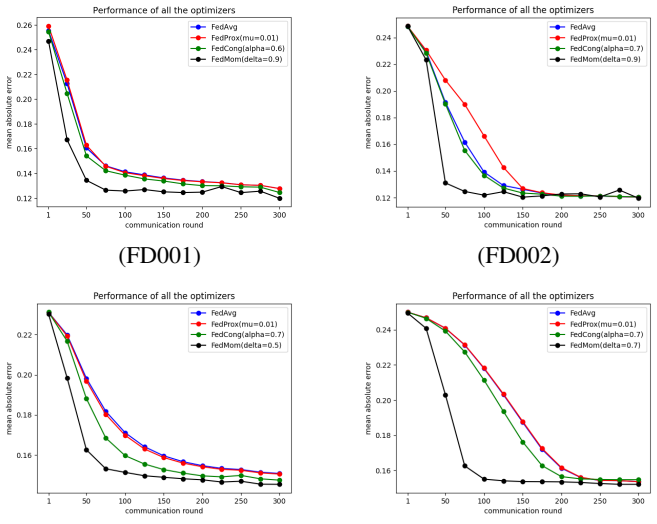


Figure 1: Performance comparison between FedAvg, FedProx, FedCong and FedMom

4 Conclusion

The main contribution of this work are the proposal and development of two new Federated Learning algorithms, namely FedCong and FedMom, and comparison with state of the art. These methods have the objective of improving the convergence speed of the Federated Learning models.

Although FedCong and FedMom greatly increased the convergence speed of the models, some limitations should be considered. Firstly, in FedCong, it is of most importance to tune the control term, α, with respect to the data distribution. If the α value is low and the current CR has a bad error representation, the global model will have difficulties to converge.

As for FedMom, the momentum parameter δ has to be tuned in order for the model to converge. During a model's update that has a poor representation of the global error, a large momentum term can cause the model to diverge even further and have constant fluctuations. This can cause difficulties in the convergence of the model.

For future work it is suggested to improve the FedCong algorithm by taking into consideration more than just the direction of the gradient descend step. For example, taking into consideration also the size of the step may help reducing the fluctuation of the global model.

In addition, it is suggested to compare the FedMom algorithm to the work proposed by Huo et al. [2] where the authors use the current model's update to estimate the new weight value, instead of using an exponentially weighted average.

References

[1] Daniel Ramage Seth Hampson Blaise Aguera y Arcass H.Brendan McMahan, Eider Moore. Communication-efficient learning of deep networks from decentralized data. In *In AAAI Fall Symposium*.

[2] Zhouyuan Huo, Qian Yang, Bin Gu, and Lawrence Carin. Heng Huang. Faster on-device training using new federated momentum algorithm, 2020.

[3] Ning Qian. On the momentum term in gradient descent learning algorithms. *Neural Networks*.

[4] Anit Kumar Sahu, Tian Li, Maziar Sanjabi, Manzil Zaheer, Ameet Talwalkar, and Virginia Smith. On the convergence of federated optimization in heterogeneous networks.

[5] Tianbao Yang, Qihang Lin, and Zhe Li. Unified convergence analysis of stochastic momentum methods for convex and non-convex optimization. *arXiv: Optimization and Control*.

Optimal lag selection for covariates in INGARCH models: an application to the analysis of air quality effect on daily respiratory hospital admissions

Ana Martins[1]
a.r.martins@ua.pt
Manuel Scotto[2]
manuel.scotto@tecnico.ulisboa.pt
Sónia Gouveia[1]
sonia.gouveia@ua.pt

[1] Institute of Electronics and Informatics Engineering of Aveiro
University of Aveiro, PT

[2] Center for Computational and Stochastic Mathematics and Department of Mathematics, IST
University of Lisbon, PT

Abstract

A comparison between strategies aiming at optimal lag selection for covariates in INGARCH models, in the context of the analysis of the association between air quality and daily number of respiratory hospital admissions in Portugal is presented. To this end, a block-forward (BF) approach is developed for automatic selection of covariates. Then, two strategies are used for optimal lag selection: (i) fixed lag (FL) approach, with optimal lag being selected as the one which maximises the cross-correlation between the covariate and the daily admissions; and (ii) changeable lag (CL) approach, with optimal lag being selected as that minimising the AIC criterion among all candidate lags. Results show that CL models have more significant covariates and lower AIC values than FL models. The coefficients of covariates simultaneously in FL and CL models are similar, despite having different optimal lags. Hence, the lag selection strategy has an impact on model fitting, which cannot be neglected.

1 Introduction

This study considers the INteger Generalised AutoRegressive Conditional Heteroskedastic (INGARCH) processes to model the association between hospital admissions and air quality. These have an ARMA-like structure, though the data generating mechanism is analogous to that of a GARCH model in the sense that the conditional mean recursively depends on the past conditional means and on the past observations [2, 3]. The INGARCH formulation incorporates link/transformation functions [8], to deal with negative serial correlation [4] and, time-varying covariates [5].
Model construction with covariates demands optimal criteria for covariate selection. The importance of such criteria is evident, as model performance can be improved by ignoring irrelevant covariates and, by considering the relevant covariates at optimal lags. These criteria should also address collinearity, as a strong association among covariates may obscure their relationship with the response and may lead to computational instability in model estimation. This paper introduces a novel method for optimal selection of time-varying covariates - the block-forward (BF). Briefly, covariates expected to induce the same effect on the response are included in one block. For each block, the significant covariate leading to the lowest Akaike Information Criteria (AIC) model is included in the model, as long as the covariates already in the model remain significant.
In time series context, the association between a response and a predictor are usually lagged. As an example, it is well-known that the maximal association between air pollution and hospital admissions may be delayed up to 7 days [7]. Traditionally, the optimal response/predictor lag is evaluated from the absolute cross-correlation function (CCF), previously to model construction. However, this procedure does not consider the possible associations among covariates. Thus, optimal lag choice in the process of covariate selection (and not *a priori*) is a promising approach, as different lagged versions of the same predictor can be thought of as a block of collinear covariates. Thus, we aim at the comparison of two strategies for lag selection, one based on the traditional CCF criterion (fixed lag, FL) and another considering the optimal AIC criterion among several candidate lags (changeable lag, CL), using the BF covariate selection method.

2 Data & Methods

2.1 Data

This study included the analysis of time series of air quality ($PM_{2.5}$, PM_{10}, NO_x, NO_2, CO, O_3 and SO_2), of temperature and of daily counts of hospital admissions (due to respiratory causes) during the 2005-2017 period. Figure 1 shows an example of a hospital admission time series, which clearly exhibits an annual seasonal pattern.
The spatial matching of air quality, temperature and hospital admissions was based on a 20km influence circumference centered around each air quality monitoring station. Hourly air quality data at 58 monitoring stations were downloaded from QualAr (`www.qualar.apambiente.pt`). Hourly temperature data at 23 spatial locations were made available by Instituto Português do Mar e da Atmosfera (`https://www.ipma.pt/`). Daily series were obtained from the maximum daily values, when at least 75% of hourly observations were available at a given day, otherwise were obtained through 1-NN imputation. Temperature series were matched to each spatial location based on their geographical proximity (measured with euclidean distance). All hospital admissions episodes registered in Portugal (2005-2017) were provided by Administração Central do Sistema de Saúde (`http://www.acss.min-saude.pt`). For each spatial location, the daily number of hospital admissions due to respiratory causes was recorded as the count of episodes connected with respiratory system diseases' (ICD-9:460–519 and ICD-10:J00-J99) from patients with address within the 20km influence circumference.

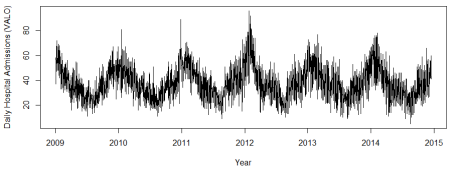


Figure 1: Hospital Admission time series at Valongo, Portugal.

2.2 INGARCH models

The INGARCH process (Y_t) assumes that the conditional distribution of Y_t is Negative Binomial i.e.,

$$Y_t|\mathcal{F}_{t-1} \sim NB(\lambda_t, \phi), \tag{1}$$

where $\lambda_t := E(Y_t|\mathcal{F}_{t-1})$ and $\phi \in (0,\infty)$ represents the dispersion parameter. Note that $Var(Y_t|\mathcal{F}_{t-1}) = \lambda_t + \lambda_t^2/\phi$ so, the limiting case $\phi \to \infty$ is the Poisson distribution with parameter λ_t. In this formulation,

$$\mathcal{F}_{t-1} := \sigma(\, Y_s, \boldsymbol{X}_{s+1},\ s \le t-1\,) \tag{2}$$

expresses the joint history of the process (up to time $t-1$) and covariates (up to and including time t). Also, the conditional expectation λ_t satisfies the recursion

$$g(\lambda_t) = \beta_0 + \sum_{k=1}^{p} \beta_k\, \tilde{g}(Y_{t-k}) + \sum_{\ell=1}^{q} \alpha_\ell\, g(\lambda_{t-\ell}) + \boldsymbol{\eta}^T\, \boldsymbol{X}_t, \tag{3}$$

where p and q are the INGARCH model orders, $\beta_0 > 0, \beta_k \ge 0$, $\alpha_\ell \ge 0$, $\forall_{k,\ell}$ and $\sum_{k=1}^{p} \beta_k + \sum_{\ell=1}^{q} \alpha_\ell < 1$. The latter condition ensures the stationary of the INGARCH process. Also, the link function $g : \mathbb{R}^+ \to \mathbb{R}$ and the transformation function $\tilde{g} : \mathbb{N}_0 \to \mathbb{R}$ were set as the natural logarithm function, to easily accommodate covariates into the model [5]. Finally, $\boldsymbol{X}_t = (X_{t,1}, \ldots, X_{t,r})^T$ is a time-varying r-dimensional covariate vector for each time t and $\boldsymbol{\eta} := (\eta_1, ..., \eta_r)^T$ is the parameter vector of the covariates coefficients. The estimation of INGARCH coefficients require a fixed order p and q. Optimal (p,q) pairs were chosen by AIC minimisation, varying from 0 to 7 in order to accommodate several INGARCH-like structures and include terms related with the presence of weekly seasonality.

2.3 Block-Forward and optimal lag selection for covariates

The block-forward (BF) selection method allows the automatic selection of significant covariates in $\boldsymbol{X}_t$. In the conventional forward method, e.g. used in linear regression, covariates are sequentially added to the model according to their statistical significance. In the BF method, the covariates are organised in blocks, where each block includes the covariates that are expected to induce a similar effect on Y_t. Consequently, the covariates in the same block are also expected to be correlated. For each block, the significant covariate leading to the lowest AIC model enters the model, as long as the other covariates remain significant (at 5% significance level). The order of the blocks is presented in Fig. 2 and reflects the current knowledge on the effect of temperature and air pollutants on hospital admissions [1]. In the BF implementation, two approaches were considered in the computation of the optimal lag between each covariate and Y_t. The fixed lag (FL) approach considers the covariate lag as that maximising the absolute values of the sample cross-correlation between the covariate and Y_t. And, the changeable lag (CL) approach that selects the optimal lag in which the BF conditions for a covariate to enter the INGARCH model are optimised. In practice, the implementation of FL and CL approaches is quite similar: while the FL approach considers the same number of candidates and covariates in one block, the CL approach considers that the number of candidates in one block is equal to the number of covariates in that block times the number of lags to be tested (in this case 8, from 0 to 7). Taking the example of block 2, FL approach tests up to 2 candidates to enter the model while CL approach will test up to 16 candidates. Note that, in both approaches, one candidate per block is selected at most.

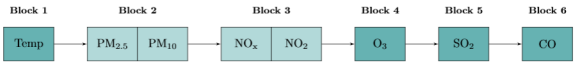


Figure 2: Blocks of covariates in the block-forward approach.

3 Results

The constructed INGARCH models were compared with respect to the number of selected covariates, the corresponding coefficients estimates and the chosen lags. Figure 3 shows the number of selected covariates out of the available for both approaches. Overall, CL models select more covariates than FL models. As an instance, temperature is selected in 54/58 CL models compared to 41/58 in FL models. Also, air quality covariates are more often selected in CL than in FL models. The median number of covariates included in the FL and CL models is, respectively, 2 and 3 covariates. Overall, both approaches show that air quality covariates are significantly associated with daily hospital admissions, beyond the well-established effect of temperature [6].

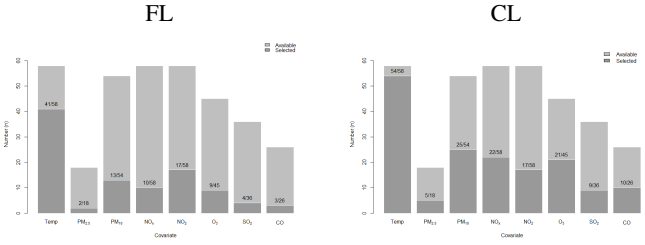


Figure 3: Barplot of the number of selected (dark grey) over the number of available (light grey) covariates for the 58 spatial locations analysed.

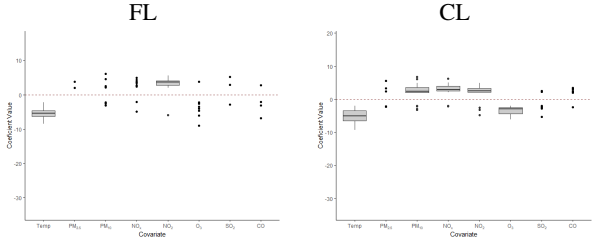


Figure 4: Distribution of the scaled coefficients at the 58 spatial locations. Boxplots are shown when there are at least 15 locations.

Lag	Temp		PM_{10}		NO_2		O_3		CO	
	FL	CL	FL	CL	FL	CL	FL	CL	FL	CL
0	0.0	27.8	10.0	8.0	68.8	5.9	10.0	19.0	0.0	0.0
1	0.0	5.6	10.0	12.0	0.0	11.8	10.0	19.0	0.0	0.0
2	0.0	5.6	20.0	24.0	0.0	17.6	0.0	9.5	0.0	20.0
3	0.0	14.8	0.0	20.0	0.0	17.6	0.0	23.8	0.0	20.0
4	4.9	13.0	20.0	16.0	0.0	5.9	0.0	9.5	0.0	30.0
5	12.2	13.0	30.0	8.0	0.0	0.0	0.0	4.8	0.0	10.0
6	4.9	11.1	10.0	0.0	12.5	5.9	0.0	9.5	50.0	10.0
7	78.0	9.3	0.0	12.0	18.8	35.3	80.0	4.8	50.0	10.0
Total (%)	100	100	100	100	100.0	100	100	100	100	100
Total (N)	41	54	10	25	16	17	10	21	4	10

Table 1: Distribution of the chosen lags according to FL and CL approach.

Figure 4 displays the distribution of the estimated scaled coefficients (i.e. coefficient divided by its standard error) for each covariate. Temperature and O_3 are negatively associated with respiratory hospital admissions, whereas the remaining air pollutants are, in general, positively associated. The magnitude of the coefficients and, the overall direction of the association are similar for both approaches. Hence, there is no major impact on the quantification of the covariate effect between approaches.
Table 1 shows the distribution of the chosen lags for some of the covariates analysed (Temp, PM_{10}, NO_2, O_3 and CO) according to each approach. There is some variability in the proportion of selected lags depending on the approach. For instance, while lag 7 is the preferred for Temp (78.0%) in the FL approach, lag 0 (27.8%) and lag 3 (14.8%) are the most frequently chosen in the CL approach. It is worthy to mention that CL models have, on average, lower AIC (< 20 units). Recall that the AIC criterion is a trade-off between information and number of covariates in a model (where increased number of covariates is penalised). Thus, the information of the covariates added pay-off the increase in complexity.

4 Conclusion

Despite the CL approach choosing more variables and different lags, the coefficients estimates remain similar for the covariates between approaches. However, the AIC of CL models is lower than that of FL models, indicating that the amount of information introduced by the additional variables in CL models pays-off the increased number of variables. Thus, tuning the lag during covariate selection is more advantageous as it increases the model performance. The trade-off is that the CL approach is computationally more demanding as both the covariates and their lagged versions are tested in the BF algorithm, which is an important aspect to consider when performing such analysis. Either way, an adequate modelling strategy is essential to assist in hospital planning and resources management and, ultimately, to contribute to better health/environmental policies.

References

[1] N. M. Ab, A. N. Aizuddin, and R. Hod. Effect of air pollution and hospital admission: a systematic review. *Annals of Global Health*, 84 (4):670, 2018.

[2] R. Ferland, A. Latour, and D. Oraichi. Integer-valued GARCH process. *Journal of Time Series Analysis*, 27(6):923–942, 2006.

[3] A. Heinen. Modelling time series count data: an autoregressive conditional Poisson model. *Social Science Research Network*, 2003.

[4] M. Ispány, V. A. Reisen, G. C. Franco, et al. On generalized additive models with dependent time series covariates. In *International Work-Conference on Time Series Analysis*, pages 289–308. Springer, 2017.

[5] T. Liboschik, K. Fokianos, and R. Fried. tscount: An R Package for Analysis of Count Time Series Following Generalized Linear Models. *Journal of Statistical Software*, 82(5):1–51, 2017.

[6] E. Martínez-Solanas and X. Basagaña. Temporal changes in the effects of ambient temperatures on hospital admissions in spain. *PloS one*, 14(6):e0218262, 2019.

[7] A. Slama, A. Śliwczyński, J. Woźnica, et al. Impact of air pollution on hospital admissions with a focus on respiratory diseases: a time-series multi-city analysis. *Environmental Science and Pollution Research*, 26(17):16998–17009, 2019.

[8] D. Tjøstheim. Some recent theory for autoregressive count time series. *Test*, 21(3):413–438, 2012.

Author Index