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.

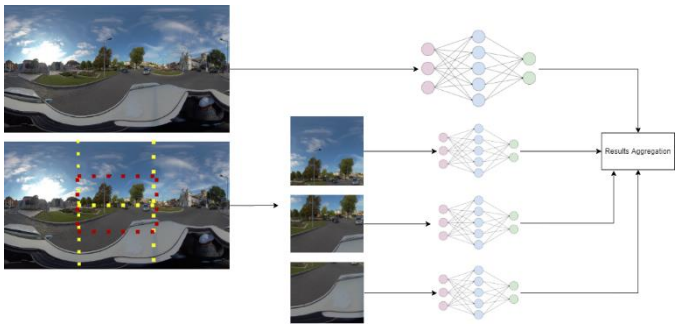

Figure 2: Proposed framework architecture with two stages. Results are aggregated with predictions from mid-region sub-divisions.

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.