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 pre-processing. 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.

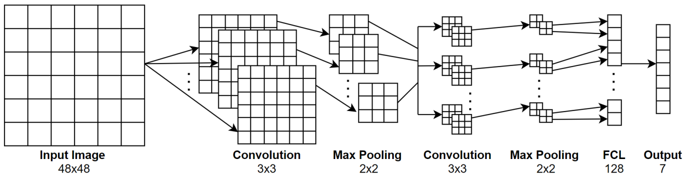

Figure 1: Proposed CNN architecture for facial expression recognition.

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.

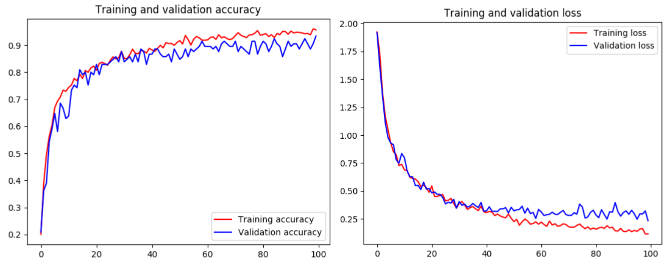

Figure 2: Training step results of the facial expression recognition model.

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:

$$A_{min} = |K_{Ymax} - E_{Ymin}| \times |S_{Xmax} - S_{Xmin}| \tag{1}$$

$$A_{current} = |K_{Ymax} - K_{Ymin}| \times |K_{Xmax} - K_{Xmin}| \tag{2}$$

$$A_{ratio} = \frac{A_{current}}{A_{min}} \tag{3}$$

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.

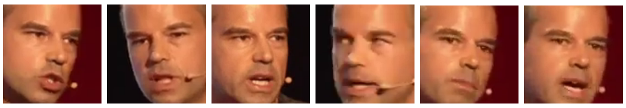

Figure 3: False positives of the facial expression recognition model. From left to right: anger, disgust, fear, happiness, sadness and surprise.

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.

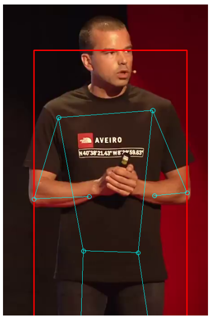

Figure 4: Processed frame from the TEDxAveiro speech segment.

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:

$$\textbf{Mood} = \frac{Expansiveness + (Negative + Neutral \times 3 + Positive \times 5)}{2} \tag{4}$$

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.