Emotion recognition using Texture Maps and Convolutional Neural Networks

Abstract

In this paper, we present a method to recognize facial expressions in video sequences considering all face movements and head behaviour. Therefore, we generate texture maps to encode these information. Next, we applied CNN models in the classification stage. Experiments on the Extended Cohn-Kanade (CK+) dataset have prove the viability of our proposal overcoming methods that analyze a single image.

1 Introduction

Facial expressions are a form of nonverbal communication which provides and convey information about the emotional state of a person. This information helps us to understand the intentions of other people, such as happiness, anger, sadness, fear, disgust, surprise, among others [Ko, 2018]. Currently, automatic facial expression has become an active research area, due to the several advances of human-computer interaction, security, and academic research [Lucey et al., 2010]. To guarantee a robust recognition of human emotional states, they must be interpreted, processed, and analyzed. Therefore, facial expressions can be described as combinations of the facial behavior, and motions performed by a human. Friesen and Ekman [1978] developed the Facial Action Coding System (FACS), which taxonomizes human facial movements by their appearance on the face. Tian et al. [2001], Bartlett et al. [2006] used the FACS to analyze and recognize the changes of facial features. In the literature, authors select the last frames of each image sequence with peak expression in their experiments without considering the head behavior and motion performed in social communication [Mollahosseini et al., 2016, Ding et al., 2017a, Zeng et al., 2018]. Similarly, the facial expression begins at the neutral frame and ends at the peak expression frame with all face movements providing additional information that improves the recognition task. Therefore, the current study considers both information to process facial expressions in videos; each video starts with a neutral expression switching to a specific expression. Thus, texture maps were generated to encode the face variations and motion until producing a specific facial expression. Experiments on The Extended Cohn-Kanade Dataset (CK+) have demonstrated the viability of our proposal overcoming methods that analyze a single image.

2 Proposed method

The pipeline of our proposed method is shown in Fig. 1. We adapted the method proposed by Ding et al. [2017b] to generate the texture maps. We first compute face landmarks (68 total points) in all frames from a video following [Nirkin et al., 2018]. Next, to describe the local facial changes, we group landmarks into three regions with 25 points, R1 (eyes, brows, and root of the nose), R2 (mouth and nasal base), and R3 (mouth and mandible). There are three types of features extracted from all combination of points. We compute for each region: *a)* the **point – point distances** between two points, resulting in $C_{25}^{2} \times N = 300 \times N$ dimensional PoP feature vector, where N is the frame number; *b)* the **point–line distances** between a point and a line formed by two adjacent points (we considered 20 lines by region), resulting in a $460 \times N$ dimensional PoL feature vector; *c)* the **line–line angles** formed between two lines in a region, obtaining a $190 \times N$ dimensional LoL feature vector. Likewise, we use RGB color images to encode the spatial feature vectors to capture temporal

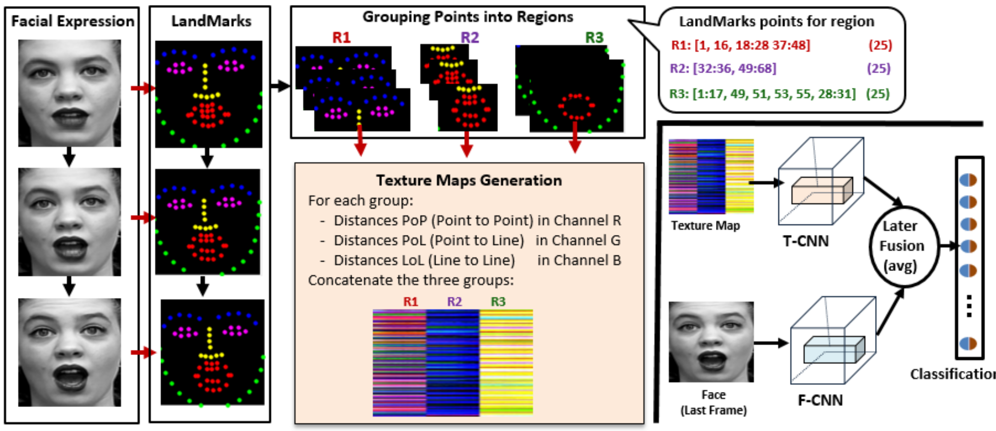

Figure 1: Pipeline of our proposed model.

information. Each column in the image represents spatial features in a frame, and each row represents the sequence of a specific feature. For each region, we resize the PoP, PoL and LoL vectors to $224 \times N$ using a bilinear interpolation. Then, we concatenate these feature vectors considering PoP in the R channel, PoL in the G channel and LoL in the B channel. Finally, we combine the texture maps from each region (R1, R2, R3) to generate a single texture map, as shown in Fig. 1.

In the classification step, we use transfer learning, *i.e.*, a pre-trained CNN model (specifically, the *imagenet-vgg-f*) [Chatfield et al., 2014] to training two ConvNets: *a)* T–CNN model, using as input the texture map to obtain spatial features; *b)* F–CNN model, using as input the last frame of a facial expression sequence to obtain local features from the face. Lastly, the final score represents the fused output scores of the two ConvNets using the average operator.

3 Experimental results and discussions

We use the extended Cohn–Kanade (CK+) dataset [Lucey et al., 2010] to evaluate the proposed framework. The CK+ consists of 593 image sequences from 123 subjects to performance eight basic facial expression categories (listed in Table 1). To conduct experiments, we follow the same experimental protocol from [Ding et al., 2017a], *i.e.*, we apply 10 fold cross-validation for training and testing. In Table 1, we compare our approach with three state-of-the-art methods in terms of average accuracy. The later fusion of T–CNN and F–CNN (T–F CNN) significantly outperform all others, achieving 96.8%. For each class, we achieve 100% of accuracy except the *sadness* emotion (75%) due to its high similarity with *anger* class. Analyzing the results, we observe that only using the T-CNN model without local information from the face achieve a score of 88.4%, due to the similarity of texture maps between different classes. Similarly, when training a single image for facial expression recognition (F–CNN model), the lack of temporary information also generates confusion in the recognition stage. Therefore, we conclude that it is necessary to combine both information to produce a robust method. Thus, in this work, we prove that using texture maps is a feasible way to encode the temporal information. As future work, we pretend to use other CNN models to improve the results achieved and testing our method on a dataset that has facial expressions with head movements (such as *affirmative* or *negative* answers).

Table 1: Comparison with the state-of-the-art methods on the CK+

Method	**Facial Expression**								**Acc**
	Anger	**Contempt**	**Disgust**	**Fear**	**Happy**	**Sad**	**Surprise**	**Neutral**	
FN2EN [Ding et al., 2017a]	99.3	90.4	100.0	100.0	97.7	94.8	98.0	94.7	**96.8**
DSAE [Zeng et al., 2018]	86.1	75.0	92.4	78.0	97.8	76.8	96.9	91.4	**89.8**
AUDN [Liu et al., 2013]	81.5	77.8	95.5	82.7	99.5	71.4	97.6	95.4	**92.1**
T–CNN (Our)	67.0	100.0	90.0	100.0	100.0	50.0	100.0	100.0	**88.4**
F–CNN (Our)	100.0	100.0	100.0	50.0	100.0	75.0	100.0	85.0	**88.8**
T–F CNN (Our)	100.0	100.0	100.0	100.0	100.0	75.0	100.0	100.0	**96.8**

References

Byoung Ko. A brief review of facial emotion recognition based on visual information. *sensors*, 18 (2):401, 2018.

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.

E Friesen and Paul Ekman. Facial action coding system: a technique for the measurement of facial movement. *Palo Alto*, 3, 1978.

Y-I Tian, Takeo Kanade, and Jeffrey F Cohn. Recognizing action units for facial expression analysis. *IEEE Transactions on pattern analysis and machine intelligence*, 23(2):97–115, 2001.

Marian Stewart Bartlett, Gwen Littlewort, Mark G Frank, Claudia Lainscsek, Ian R Fasel, Javier R Movellan, et al. Automatic recognition of facial actions in spontaneous expressions. *Journal of multimedia*, 1(6):22–35, 2006.

Ali Mollahosseini, David Chan, and Mohammad H Mahoor. Going deeper in facial expression recognition using deep neural networks. In *2016 IEEE Winter conference on applications of computer vision (WACV)*, pages 1–10. IEEE, 2016.

Hui Ding, Shaohua Kevin Zhou, and Rama Chellappa. Facenet2expnet: Regularizing a deep face recognition net for expression recognition. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, pages 118–126. IEEE, 2017a.

Nianyin Zeng, Hong Zhang, Baoye Song, Weibo Liu, Yurong Li, and Abdullah M Dobaie. Facial expression recognition via learning deep sparse autoencoders. *Neurocomputing*, 273:643–649, 2018.

Zewei Ding, Pichao Wang, Philip O Ogunbona, and Wanqing Li. Investigation of different skeleton features for cnn-based 3d action recognition. In *2017 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, pages 617–622. IEEE, 2017b.

Yuval Nirkin, Iacopo Masi, Anh Tran Tuan, Tal Hassner, and Gerard Medioni. On face segmentation, face swapping, and face perception. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 98–105. IEEE, 2018.

Ken Chatfield, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Return of the devil in the details: Delving deep into convolutional nets. *arXiv preprint arXiv:1405.3531*, 2014.

Mengyi Liu, Shaoxin Li, Shiguang Shan, and Xilin Chen. Au-aware deep networks for facial expression recognition. In *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pages 1–6. IEEE, 2013.