Recognition of Facial Expressions Based on Deep Conspicuous Net

João Paulo Canário(✉) and Luciano Oliveira

Intelligent Vision Research Lab, Federal University of Bahia,
UFBA, Salvador, Brazil
jopacanario@gmail.com, lreboucas@ufba.br
http://www.ivisionlab.eng.ufba.br/

Abstract. Facial expression has an important role in human interaction and non-verbal communication. Hence more and more applications, which automatically detect facial expressions, start to be pervasive in various fields, such as education, entertainment, psychology, human-computer interaction, behavior monitoring, just to cite a few. In this paper, we present a new approach for facial expression recognition using a so-called deep conspicuous neural network. The proposed method builds a conspicuous map of region faces, training it via a deep network. Experimental results achieved an average accuracy of 90% over the extended Cohn-Kanade data set for seven basic expressions, demonstrating the best performance against four state-of-the-art methods.

Keywords: Conspicuity · Facial expression · Deep learning

1 Introduction

Non-verbal language can be highlighted as one of the first forms of human communication, and consequently a source of countless studies in science. Particularly, facial expression is one of the most powerful, immediate and natural non-verbal ways that humans can use to transmit their emotions and intentions [2]. Also, face is able to express emotions so soon as a person can speak or perceive his/her feelings [5].

Adopting Charles Darwin's starting premise [1], which stated that the mammals understand and show their emotions from a set of facial expressions, Ekman and Friesen [4] initially suggested that there exist six primary emotions plus neutral (e.g, happy, sadness, fear, disgust, surprise and anger), with each one of them having singular and universal facial expressions and characteristics; later, they have also included contempt as a primary expression [8].

In contrast with facial expression recognition, emotion recognition is a pure interpretation of the expression, and it frequently demands a comprehension of a given situation, along with the evaluation of all contextual information surrounding [3]. With that in mind, some research areas, such as affective computing, try to give to computers the ability to recognize and feel emotions. In the near

A. Pardo and J. Kittler (Eds.): CIARP 2015, LNCS 9423, pp. 1–8, 2015.
DOI: 10.1007/978-3-319-25751-8_31

future, it may be possible to give more particular, natural and proper guidance to end-users in the human-computer interaction [6].

Nowadays, starting from facial expression detection to ultimately accomplish emotion recognition, some works have been achieving high accuracy. Chew et al. [9] explored a person-independent system using constrained local models (CLM) to highlight the face shape and recognize facial expression by using local binary pattern (LBP) descriptor with an SVM classifier. Lee and Chellappa [10] introduced a framework for facial motion modeling; they represented the faces as a sparse localized motion dictionaries obtained by a motion flow estimation, which were then classified by an nearest neighborhood (NN) classifier. Nie, Wang and Ji [11] have dealt with facial expression recognition by means of a type of classification problem over multi-dimensional sequence data; they extracted spatio-temporal patterns in high-dimensional motion data using an improved restricted Boltzmann machine (RBM), where pairwise potential energy functions were used; the main goal was to break the assumption of the input data dependence by means of a standard RBM model. Shojaeilangari et al. [12] introduced a histogram of local phase and local orientation of gradients achieved from a sequence of face images, as a descriptor of facial expressions.

Differently from the other works, we propose a method that combine conspicuous maps representation (addressed in Section 2.1), and a deep learning approach to classify that conspicuous regions, as described in Section 2.2. The conspicuous maps highlight the most salient areas of the face, avoiding the classification of unnecessary areas of the face that does not influence on the facial expression (*e.g.*, ears, top of the head and hairs). The convolutional neural network (CNN) helps to learn different high-level features (over the eyes, mouth and nose). This is so, since the deep net can locally sharing weights at high layers providing a greater abstraction power. To assess the performance of the proposed method, the extended Cohn-Kanade (CK+) data set [17] was used as a referential comparison. Over that data set, our method achieved the best performance when compared against four state-of-the-art works [9] [10] [11] [12], reaching an average accuracy of 90%, considering all the expressions.

This paper is structured as follows: In Section 2, our proposed method is described. Section 3 presents the experimental results. Finally, Section 4 draws the conclusions, as well as, suggestions for future works.

2 On Deeply Learning Facial Expressions

2.1 Conspicuity Maps

An object is more usually noticed in a scene based on their behavioral relevance. In case of the facial expressions, the regions that are capable to draw more visual attention are in the t-region of the face (eyes and nose), which ultimately highlight most of the facial movements to be analyzed. To detect the salient regions of the face, similarly to the way to achieve the intense maps in [18], the input image I is progressively downsampled using a Gaussian pyramid [19], which consists of low-pass filtering and sub-sampling versions of the input image, in eight octaves σ. From the Gaussian pyramid, each feature is computed by a set of linear "center-surround" differences, denoted as $\ominus$, and they are implemented in the model, as the difference between fine and coarse scales, whose center is a pixel at scale $c \in \{2, 3, 4\}$, and the surrounding is the corresponding pixel at scale $s = c + \delta$, where $\delta \in \{3, 4\}$. The referred differences are computed in a set of six maps, given by

$$\mathcal{I}(c, s) = \mid I(c) \ominus I(s) \mid \ . \tag{1}$$

The conspicuity map $\overline{\mathcal{I}}$ is obtained by a cross-scale addition, "$\bigoplus$", which works by reducing each map to scale 4 and point-by-point addition, according

$$\overline{\mathcal{I}} = \bigoplus_{c=2}^{4} \bigoplus_{s=c+3}^{c+4} \mathcal{N}(\mathcal{I}(c, s))\,. \tag{2}$$

A normalizing operator $\mathcal{N}(.)$ is used to globally promotes maps where a small number of strong peaks of activity – conspicuous locations – is presented, while globally suppresses maps, which contain numerous comparable peak responses. The normalizing operation consists of: (i) Normalizing the values in the map to a fixed range *[0..M]*, in order to eliminate modality-dependent amplitude differences; (ii) finding the location of the map's global maximum M and computing the average $\overline{m}$ of all its other local maxima; and (iii) globally multiplying the map by $(M - m^2)$. After that, a fixed threshold defined empirically is applied in the map in order to create a binary mask; this binary image finally highlight the salient regions (refer to Fig. 1 for visual examples of the method steps).

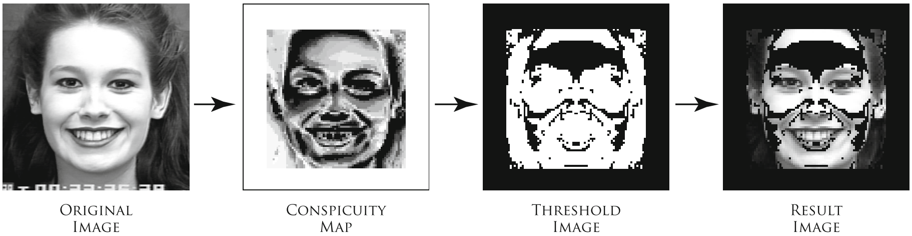


Fig. 1. Generation of the conspicuity regions. Left to right: First image is the original image after alignment, cropping and color normalization. Second image is the generated conspicuity map of the face. Third image is the thresholded image and, finally, the fourth image is the image only with the most salient facial regions.

2.2 Deep Learning

Deep learning can be described as a learning experience at various levels of representation, corresponding to different levels of abstractions. To consider a neural network deep, it is necessary that the input of the deep network pass through several non-linearity filters before being output.

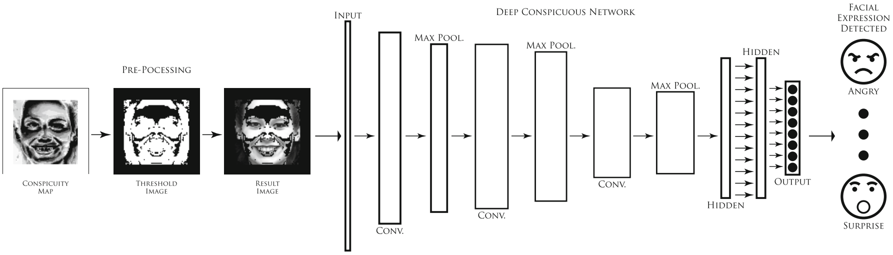


Fig. 2. General architecture of the proposed method. Left to right: Conspicuous map representation and the architecture of the deep CNN.

In this paper, we focus on combining a strategy of a salience map and a deep convolution network. The proposed method is based on a coarse-to-fine approach, by considering the conspicuous maps, the three convolution layers followed by max pooling, two hidden layers fully connected and one output layer, also fully connected, with eight possible outputs, as illustrated on Fig. 2.

Some important aspects of the network architecture have been taken into consideration. First, predicting an expression valence from large input regions is a high-level task. So, because deeper structures help to form high-level information, our convolution networks should be deep, as well. Second, since deep structures tend to be very hard to train, and to obtain performance improvement, the network should locally share weights of neurons. On the other hand, globally sharing weights does not work well on images with fixed spatial layout, such as faces; once eyes and mouth may share low-level features, they are very different at high-level. This way, for networks whose inputs contain different semantic regions, locally sharing weights at high layers is usually more effective for learning different high-level features. The idea of locally sharing weights was originally proposed by Huang, Lee and Learned-Miller [16].

Considering the proposed architecture, the input layer of the deep neural network is denoted by a vector of size hxw where h and w is, respectively, the height and width of the input image $I(h,w)$. Also, the input is 2D since color information is not used. Convolutionals layers are denoted by $C(f,\ s)$, where f is the number of the square convolution kernels, or filters, and s is the size of the filters. Each map in the convolutional layer is evenly divided into p by q regions, and weights are locally shared in each region. Pooling layer is denoted as $P(ds)$. The parameter ds is the size of the square pooling regions, which are not overlapped. The fully connected layers are denoted as $F(n)$, where n is the number of nodes at the current layer.

2.3 Deep Conpiscuous Neural Net

Before training the convolutional neural network with the data set images, all images were pre-processed. The preprocessing module in Fig. 1 illustrates all steps described below

1. First all the faces images were aligned and cropped to 380x380 pixel wide.
2. Second image normalization was made with a contrast filter and grey-level transformation for those images that were not already in grey-level, yet.
3. Finally, the salient regions of face are obtained by computing conspicuity maps, described in Section 2.1, over each training image.

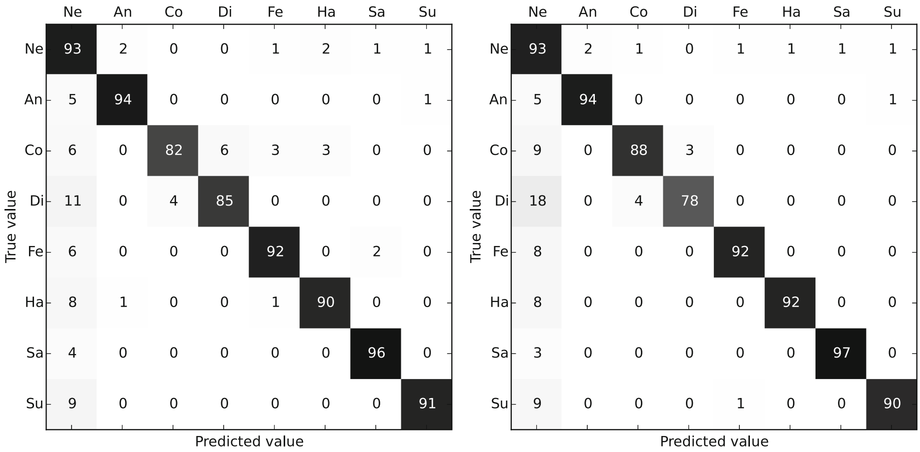
After the preprocessing phase, the images with salient regions are used to train the deep neural network, and for training validation a stratified K-Folds cross, where the number of folds is equals to 3, is used.

In our work, we evaluate two different deep neural network architectures: (i) The first one is so-called Deep conspicuous network (DCN), built on three convolutional layers and two fully connected layers; each convolutional layer is followed by a 2x2 max-pooling layer; starting with 32 convolutional filters, this number is doubled with every convolutional layer, which has 3x3 and 2x2 filters. The fully connected hidden layers have 500 units and a output layer is a full connected layer with eight possible outputs (one for each detected expression plus neutral). Figure 2 depicts an overview of our system using the DCN approach. (ii) The second architecture evolved from this first one, and it was coined as dropout DDCN (DDCN); in that second architecture, we increased the number of units of the fully connected hidden layers from 500 to 1000 units; also the learning rate and momentum overtime were changed during the training, after Sutskever et al. [14]; dropout layers were added between the existing layers, assigning dropout probabilities to each one of them. Dropout is a popular regularization technique introduced by Hinton et al. [15] to reduce the overfitting on large neural networks.

3 Experimental Results

To assess the performance of the proposed method, we have performed experiments for emotion recognition over the widely adopted CK+ data set [17]. The CK+ facial expression data set is an extension of the original Cohn-Kanade Database [13] and consists in image sequences (frontal view) of 123 students of different ages, gender and ethnicity, performing each one the seven basic facial expressions: Anger – An, Contempt – Co, Disgust – Di, Fear – Fe, Happy – Ha, Sad – Sa and Surprise – Su, plus the Ne – Neutral one. The neutral expression is obtained on the first image of each facial expression sequence and a total of 593 image sequences or 10792 separated images were generated. All images the image were considered, following the same protocol suggested in [17], and followed by all compared works. In the experiments, the CK+ data set was split into random train and test subsets. Four state-of-the-art methods were used to compare the performance of our detector.

According Fig. 3, DCN achieved the following results: Neutral - 93%; Angry 94%; Contempt 82%; Disgust 85%; Fear 92%; Happy 90%; Sadness 96%; Surprise 91%, while, DDCN performance was: Neutral - 93%, Angry 94%; Contempt 88%; Disgust 78%; Fear 92%; Happy 92%; Sadness 97%; Surprise 90%. If keep neutral expressions out, the average accuracy of all the expressions is 90%, which is 2%



	Ne	An	Co	Di	Fe	Ha	Sa	Su
Ne	93	2	0	0	1	2	1	1
An	5	94	0	0	0	0	0	1
Co	6	0	82	6	3	3	0	0
Di	11	0	4	85	0	0	0	0
Fe	6	0	0	0	92	0	2	0
Ha	8	1	0	0	1	90	0	0
Sa	4	0	0	0	0	0	96	0
Su	9	0	0	0	0	0	0	91

(a)

	Ne	An	Co	Di	Fe	Ha	Sa	Su
Ne	93	2	1	0	1	1	1	1
An	5	94	0	0	0	0	0	1
Co	9	0	88	3	0	0	0	0
Di	18	0	4	78	0	0	0	0
Fe	8	0	0	0	92	0	0	0
Ha	8	0	0	0	0	92	0	0
Sa	3	0	0	0	0	0	97	0
Su	9	0	0	0	1	0	0	90

(b)

Fig. 3. Confusion matrix of the two trained deep conspicuity facial expression architectures, Fig. (a) represents the DCN architecture, and Fig. (b) represents the DDCN architecture. The analyzed expressions were Neutral – Ne, Anger – An, Contempt – Co, Fear – Fe, Happiness – Ha, Sadness – Sa and Surprise – Su.

Table 1. Classification accuracy (%) of our method and the other four state-of-art approaches, ordered by the average performance. A bold number indicates the best performance method.

Approach	**An**	**Co**	**Di**	**Fe**	**Ha**	**Sa**	**Su**	**Avg.**
DCN	94	82	85	**92**	90	96	91	**90**
DDCN	94	**88**	78	**92**	92	**97**	89	**90**
Nie, Wang and Ji [11]	**97**	72	89	84	**100**	78	97	88
Shojaeilangari et al. [12]	90	–	**96**	66	**100**	78	**98**	88
Lee and Chellappa [10]	84	81	89	63	91	80	93	83
Chew et al. [9]	70	52	92	72	94	45	93	74

better than the best state-of-the-art method studied. In Table 1, the comparison of our model with some recent approaches are summarized. Our method reaches the best accuracy on three expressions (contempt, fear and sadness) among all the methods listed. Although our model did not achieve the best accuracy for all expressions, it does not fall behind very much the other detectors.

4 Conclusion

In this paper, we presented a new approach for facial expression recognition called deep conspicuous neural network. The proposed method achieved the best average accuracy of 90% on the CK+ data set, considering seven basic emotions, against four state-of-the-art methods. Our method relied on a salient conspicuous maps classified by a deep neural network.

For that proposed map, we created two model architectures, denoted as DCN and DDCN. Although the average performance of both architecture were the same, it is noteworthy that better results in four expressions (contempt, fear, happy, sadness) were achieved, proving the improvement of DDCN over the DCN. For future work, we are working on a pre-training approach to initialize our network with better epochs, layers and weights. Also, we are exploring the use of concatenated deep network structures that will automatically segment the most conspicuous regions of the face.

References

1. Darwin, C., Ekman, P., Prodger, P.: The Expression of the Emotions in Man and Animals. Oxford University Press, USA (1998)
2. Nakatsu, R., Nicholson, J., Tosa, N.: Emotion recognition and its application to computer agents with spontaneous interactive capabilities. Knowledge-Based Systems **13**, 497–504 (2000)
3. Fasel, B., Luettin, J.: Automatic facial expression analysis: a survey. Pattern Recognition **36**, 259–275 (2003)
4. Ekman, P., Friesen, W.V.: Measuring facial movement. Environmental psychology and nonverbal behavior. Human Sciences Press (1976)
5. Tian, Y.L., Kanade, T., Cohn, J.F.: Recognizing action units for facial expression analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence **23**, 97–115 (2001)
6. Busso, C., Deng, Z., Yildirim, S., Bulut, M., Lee, C.M., Kazemzadeh, A., Lee, S., Neumann, U., Narayanan, S.: Analysis of emotion recognition using facial expressions, speech and multimodal information. In: Proceedings of the 6th International Conference on Multimodal Interfaces, pp. 205–211. ACM (2004)
7. Ekman, P., Friesen, W.V., Ellsworth, P.: Emotion in the human face: Guidelines for research and an integration of findings. Elsevier (2013)
8. Friesen, W.V., Ekman, P.: EMFACS-7: Emotional Facial Action Coding System. Unpublished manuscript, University of California, San Francisco (1983)
9. Chew, S.W., Lucey, P., Lucey, S., Saragih, J., Cohn, J.F., Sridharan, S.: Person-independent facial expression detection using constrained local models. In: IEEE International Conference on Automatic Face & Gesture Recognition and Workshops (FG 2011), pp. 915–920 (2011)
10. Lee, C.S., Chellappa, R.: Sparse localized facial motion dictionary learning for facial expression recognition. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 3548–3552 (2014)
11. Nie, S., Wang, Z., Ji, Q.: A generative restricted boltzmann machine based method for high-dimensional motion data modeling. Computer Vision and Image Understanding (2015)

12. Shojaeilangari, S., Yau, W.Y., Li, J., Teoh, E.K.: Multi-scale Analysis of Local Phase and Local Orientation for Dynamic Facial Expression Recognition. Journal of Multimedia Theory and Application, 1 (2014)
13. Kanade, T., Cohn, J.F., Tian, Y.: Comprehensive database for facial expression analysis. In: Fourth IEEE International Conference on Automatic Face & Gesture Recognition, pp. 46–53 (2000)
14. Sutskever, I., Martens, J., Dahl, G., Hinton, G.E.: On the importance of initialization and momentum in deep learning. In: Proceedings of the 30th International Conference on Machine Learning (ICML 2013), pp. 1139–1147 (2013)
15. Hinton, G.E., Srivastava, N., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.R.: Improving neural networks by preventing co-adaptation of feature detectors. CoRR (2012)
16. Huang, G.B., Lee, H., Learned-Miller, E.: Learning hierarchical representations for face verification with convolutional deep belief networks. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 2518–2525 (2012)
17. Lucey, P., Cohn, J.F., Kanade, T., Saragih, J., Ambadar, Z., Matthews, I.: The Extended Cohn-Kanade Dataset (CK+): a complete dataset for action unit and emotion-specified expression. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 94–101 (2010)
18. Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. IEEE Transactions on Pattern Analysis & Machine Intelligence **20**, 1254–1259 (1998)
19. Greenspan, H., Belongie, S., Goodman, R., Perona, P., Rakshit, S., Anderson, C.H.: Overcomplete steerable pyramid filters and rotation invariance. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 222–228 (1994)