Emotion Identification by Facial Landmarks Dynamics Analysis

Alessandra Bandrabur, Laura Florea and Corneliu Florea
Image Processing and Analysis Laboratory,
University "Politehnica" of Bucharest, Romania
Email: abandrabur, lflorea, cflorea@imag.pub.ro

Matei Mancas
IT Department, Faculty of Engineering,
University of Mons, Belgium
Email: matei.mancas@umons.ac.be

Abstract—In this paper we concentrate our efforts on the analysis of the facial landmarks dynamics as being a relevant method to access the subject's emotion. Given the person's facial landmarks we describe their trajectory with respect to the neutral pose and out of this trajectory we extract relevant features that are subsequently entered into a classification system for the actual recognition of emotion. This procedure provides a robust estimation, that has little sensibility with respect to landmark accuracy. Moreover it brings promising results for complex emotion classes. The approach is extensively tested on the facto emotion database, namely Cohn-Kanade+. The reached performance is comparable with state of the art, but on the complete palette of possible emotions.

Keywords—*Face Landmark, Temporal dynamics, Emotion recognition.*

I. INTRODUCTION

Significant efforts have been made recently by scientists and engineers for researching multi-modal human-machine communication, like automatic speech recognition, gesture recognition, automatic human expression and emotion recognition. But all these fields of study still provide challenges. In this paper we focus on a method for emotion recognition, as it is one of the foremost non-verbal channel of human interaction. The sequences are gray-scale images of still face images.

Regarding the importance of expression recognition in practical applications we refer the reader to the report of Golomb and Sejnowski [1]. Summarizing, the recognition of face expression is the foundation for a plethora of applications such as medicine, security and computing technology like monitoring of fatigue [2], research of depression [3] or pain [4], interactive gaming, health support appliances etc. There is also theoretical interest of medical and cognitive scientists for this area [5]. A detailed review of the emotion detection methods is in the work of Zeng et al. [6].

There have been proposed various systems for description of facial expressions; yet the most used was proposed by Ekman and Friesen in 1978 [7] and reviewed in 2002 [8] and it is called FACS (Facial Action Coding System). The key concept of the FACS related literature is the universality of emotions: the basic ones are recognized equally by humans of all cultures and origins. In the initial work Ekman et al. [7] concluded that there are six basic, different facial expressions that exist, besides the neutral pose: anger, disgust, fear, happiness, sadness and surprise. The FACS theory states

that face muscles can produce a number of 46 basic facial actions named Action Units (AU) and all the fundamental expressions are produced by combinations of AUs.

The AUs are dynamic and have three phases: onset, apex and offset, and it is very difficult to determine the AUs from a single, no-reference, image. Due to this reason many state of the art systems for face expression analysis used sequences of images of the same person: a neutral image and images with the expression [9], [10], [11]. Another reason for using not only one image of the subject, but a sequence of images consists in the fact that temporal facial dynamics is crucial for the categorization of complex psychological states like various types of pain and mood [12] and also for differentiating between posed and spontaneous facial expressions [13].

One can classify the existing attempts of automatically identifying the face expressions in holistic and feature based. The first category, i.e. the holistic approaches, consider the face as a whole, indivisible. Our method belongs to the second category, the feature-based approaches. These approaches extract relevant data by performing local analysis.

The remainder of this paper is organized as follows: Section II makes a brief summary of the state of the art for feature based approaches for emotion recognition; in Section III we briefly describe the CK+ database and face fiducial points location; in Section IV our proposal for emotion recognition is presented; Section V deals with results and discussions on the here-proposed algorithm, while in the last section some conclusions are drawn.

II. PRIOR ART

A. Feature Based Approaches for Emotion Recognition

Automatic emotion recognition from facial images has been an important research topic in the last years due to the wide-range of applications. The target of facial emotion recognition systems is the classification of the image into one of the basic emotion-specific expressions. Such a system is usually composed of three stages: face detection and preprocessing, feature extraction and expression classification. The first stage deals with detecting the face in the image, aligning and normalizing it to a pre-defined grid. The second stage tackles facial feature extraction and representation, including a possible dimensionality reduction. The features can be geometrical or appearance features, or even both of them. In the third stage, the actual classification step, the system may classify each image into the basic emotion or can perform AU

classification and then use the FACS system. The systems can use static images or can also use the temporal information.

Tian et al. used geometric features for AU recognition in [14]. They searched for a set of landmark points in the AU regions and they used them for geometric feature representation. Zheng et al. [15] and Guo et al. [16] also used either the geometric positions of a set of 34 fiducial points manually located on the face, or a set of Gabor wavelet coefficients extracted from the face at the location of those points. Other popular appearance features used for emotion recognition are LBP used in [17] or SIFT used in [18].

Until recently there was a debate regarding the transition from Action units to expression and emotion. Yet, by the introduction of the Cohn-Kanade+ database [19] there has been reached a consensus about the associations.

Inspired by Valstar et al. [9] we will use the geometric positions of 44 landmarks on the face as the first descriptor and we will construct a second descriptor based on the positions of these points in a temporal sequence.

III. CK+ DATABASE

The Extended Cohn-Kanade (CK+) database [20], [19] contains 592 expression-labelled image sequences from 123 professional posers performing one of the seven discrete emotions: happiness, sadness, surprised, contempt, disgust and anger. All the sequences start with a neutral expression and end with a peak expression (apex). The relation between the posed expression and the emotion was retrieved by consensus with FACS manual and observer opinion. Each apex is annotated with one of the seven emotions.

A. Face fiducial points location

The CK+ database creators also provided a set of landmark position for each frame. These landmarks are obtained using the AAM tracker [21] initialized on manually annotated keyframes. While the accuracy of the tracker is not provided for the CK+ database, yet this is provided for a different yet similar database (UNBC - Mc Master Shoulder Pain - [22] - describes expressions in terms of action units) and it is in the performance range of automatic systems such as the one proposed by Valstar et al. [23] or by Zhu and Ramanan [24].

To evaluate the impact of landmarks, we will add specific testing that shows that very high precision of landmarks localization is not mandatory for the proposed system.

IV. PROPOSED ALGORITHM

The here-proposed method has the classical structure of a feature-based one: face detection and preprocessing, feature extraction and expression classification. The classification step is done in a late fusion scheme using two multi-layer perceptrons and a support vector machine. The schematic of the here-proposed algorithm can be seen in Fig. 1.

A. Face detection and preprocessing

The face is detected from each frame in the sequence using the classical boosted set of simple Haar classifiers proposed by Viola and Jones [25], as implemented in OpenCV. On

Fig. 2. Illustration of the landmark set used. Image taken from Cohn-Kanade+ database [19].

each face, 44 fiducial points, that can be seen in Fig. 2, are considered to be given. From those landmarks 3 are on the nose, 18 in the mouth area, 12 points are on the eyes, 10 on the eyebrows and 1 point on the chin. For testing reasons, the ground truth landmarks will be considered and different levels of noise will be added on them. Similar to other state of the art systems [10], the proposed method uses one neutral frame and 3 frames, where the emotion is in the apex phase, to recognize and classify the basic emotions.

The normalization and scaling of the face and landmarks are done with an affine transformation. It consists in applying a linear combination of translation, rotation and scaling: (1) the faces are rotated in vertical position based on the nose vertical line and the inner corner horizontal eye line, because these points are not influenced by facial muscle contractions; (2) all the faces are scaled with respect to a medium nose length, of 40 pixels (experimentally chosen); (3) the resulting Cartesian system is centered in the nose tip, by shifting all the faces.

In order to remove the error accumulation, a Gaussian filter is applied on the point coordinates from the sequence.

B. Feature extraction

Facial muscle contractions modify the shape and the position of the face fiducial points. These modifications are tracked by defining four pairs of basic geometric-features characteristics. The first pair of descriptors is based on the coordinates of each point:

$$f_1(p_i(t)) = x_i(t); f_2(p_i(t)) = y_i(t). \qquad (1)$$

where $p_i(t)$ is the fiducial point i with $i = [1; \ldots; 44]$ from the tth frame and $p_i(t)$ has the coordinates $(x_i(t), y_i(t))$.

To capture the temporal information we define the next descriptor pair, which is based on the difference between the current frame and the neutral frame from the sequence:

$$f_3(p_i(t)) = ||x_i(t) - x_i(0)||; f_4(p_i(t)) = ||y_i(t) - y_i(0)||. \quad (2)$$

where $(x_i(t), y_i(t))$ are the fiducial point coordinates at the current frame $t > 0$ and $(x_i(0), y_i(0))$ are the fiducial point coordinates at the frame presenting a neutral expression.

To track the rate of change during the facial muscle movements, the first derivative with respect to time is computed:
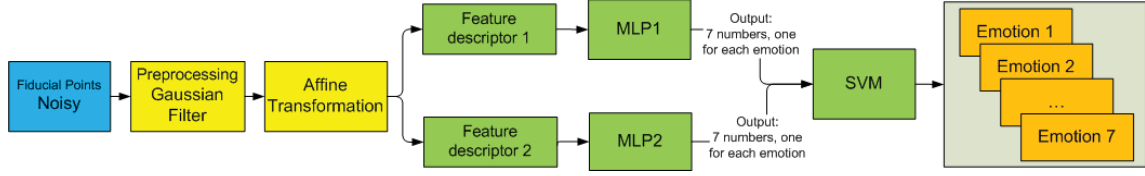
Fig. 1. The schematic of the proposed algorithm.

$$f_5(p_i(t)) = \frac{dx_i(t)}{dt}; f_6(p_i(t)) = \frac{dy_i(t)}{dt}. \qquad (3)$$

The last pair of descriptors is based on a second-order polynomial which approximates the coordinates movement. Based on the findings from [9], in order to determine the changes in neuromuscular facial action, a temporal window of seven frames is needed. Using this temporal window, we take into account the coordinates during 7 frames and we find the best movement approximation by a second-order polynomial function g. The movement of the x and respectively y coordinates of each of the fiducial points will be approximated with:

$$g(x_i(t)) = a(x_i)t^2 + b(x_i)t + c(x_i); \qquad (4)$$

$$g(y_i(t)) = a(y_i)t^2 + b(y_i)t + c(y_i). \qquad (5)$$

where t is the center frame of the seven frames window. Due to the fact that the polynomial coordinates a, b and c comprise best the movement, the resulting pair of descriptors will be:

$$f_7(p_i(t)) = [a(x_i), b(x_i), c(x_i)]; \qquad (6)$$

$$f_8(p_i(t)) = [a(y_i), b(y_i), c(y_i)]. \qquad (7)$$

Two final feature vectors are computed based on these pairs of descriptors. The first feature vector consists in the concatenation of f_1 and f_2 and has 88 values. The second feature vector, which will describe the temporal part of the expression, is formed by concatenating the pairs f_3 to f_8 resulting in a 440 feature-long vector. No dimensionality reduction is done.

A late fusion scheme is applied by combining the prediction scores of two multi-layer perceptrons (MLP) and the final score was predicted by a support-vector machine (SVM). Each of the classifiers were trained to do regression.

C. Training and testing

We perform the 8-fold crossvalidation on the training data, for comparing with the state of the art methods: we extract from the emotion sequence the first neutral frame and the last three frames (apex), as can be seen in Fig. 3. The dataset was split into 8 subsets, so that the subjects of any two subsets are not overlapped. We run the algorithm eight times, and each time we train on seven folds and test on the remaining fold. The accuracy is computed as the average of the all runs.
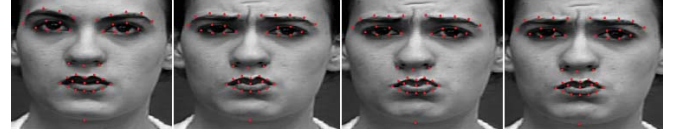

Fig. 3. Modification of landmark distribution while the face expression goes from neutral to apex (sadness). Image sequence taken from Cohn-Kanade+ database [19].

TABLE I. THE RECOGNITION RATE [%] OF THE PROPOSED ALGORITHM ON THE CK+ DATABASE, WHEN FEATURES ARE DEGRADED BY NOISE.

Uniform Noise in	Classification		
	MLP1	MLP2	SVM
[-1, 1]	83.65	92.14	90.96
[-2, 2]	83.30	91.28	90.43
[-5, 5]	81.19	84.51	86.18
[-10, 10]	75.21	70.99	76.72

To enlarge the training dataset, we add ten times uniform random noise in the range [-2,+2] pixels on all the fiducial points coordinates.

The first MLP (MLP1) is fed with the first feature vector, which contains the basic coordinates (88 size), while the second one (MLP2) is fed with the second feature vector, which covers very well the facial movements (440 size). MLP1 has two hidden layers, with 20 and 10 neurons, while MLP2 has two hidden layers with 70 and 60 neurons. Each neural network is a multi-classifier for the emotion recognition, which is a seven classes approach. Their outputs are two vectors of 7 length, which are concatenated and fed into a SVM as implemented in LibSVM [26]. We chose a RBF kernel from experimental attempts and due to the fact that the data is nonlinear. For testing we followed the same procedure.

V. RESULTS

a) Landmark localization influence: First aspect of the discussion regarding the achieved results relates to the influence of correct features. As previously mentioned we started with annotated landmarks. These landmarks are degraded by a uniform random variable and then the features $f_1, \ldots, f_8$ are computed for testing. The influence of noise over the results is presented in table I.

As one can see, small noise, from $[-1, 1]$ to $[-2, 2]$, has little impact over the recognition rate. However, expectable, as the randomness in feature increases, with a noise from $[-5, 5]$ to $[-10, 10]$, the overall recognition rate decreases.

The proposed algorithm improves the generalization, while the landmarks accuracy decreases. A single MLP leads to better results with more precisely localized facial landmarks.

TABLE II. THE RECOGNITION RATE [%] OF THE PROPOSED ALGORITHM ON THE CK+ DATABASE, COMPARED TO VARIOUS STATE OF THE ART METHODS. ACRONYM EXPLANATION LIES IN TEXT.

Method	Class number	Recognition rate
Proposed	**7**	**92.14**
PTS - [19]	7	66.68
CAPP - [19]	7	80.87
AAM - [19]	7	88.25
BDBN - [10]	6	96.7
ASR - [11]	5	97.5

b) Comparison with state of the art: A detailed comparison with state of the art methods may be followed in table II. Regarding the state of the art performance a set of three baseline methods is provided by Lucey et al. [19] in the paper introducing Extended Cohn Kanade database. They report the recognition performance for all the 7 emotions when various features are used: PTS - when only the landmarks are used, CAPP - canonical appearance (i.e. the interior of the shape formed by points) is used and when both points and appearance are put together forming the active appearance model of the face (AAM). As can one notice, although we only rely on landmarks we significantly outperform the PTS based method; yet the proposed method is superior for the other choices. However, we must stress that Lucey et al. [19] give individual decision for each frame independently, while we use a sequence of frames.

More recently, Liu et al. [10] and Mery and Bowyer [11] report results on the CK+ database. In both cases, multiple frames are used for decision. While at a first glance, the here proposed performance is lower than theirs, we must stress that they did not test for all emotions, while we do. Typically the "contempt" emotion has a lower representation in the database and recognition results on it are much weaker (see also tables 5-7 from [19]).

VI. CONCLUSIONS

In this paper we have proposed a new method to recognize emotions in sequence of images. Given the location of face fiducial points on each frame we describe the dynamics of each landmark to form features presenting the emotion case. A late fusion system was implied for actual recognition. As the CK+ database is with posed facial expressions, we consider that future research should try to estimate the actual performance on database with genuine expressions. Concluding, the future research will extend the algorithm to take into account natural environments.

VII. ACKNOWLEDGEMENTS

REFERENCES

[1] B. Golomb and T. Sejnowski, "Benefits of machine understanding of facial expressions," in *NSF Report Facial Expression Understanding*, 1997, pp. 55 – 71.

[2] Q. Yang, C. Li, and Z. Li, "Application of ftgsvm algorithm in expression recognition of fatigue driving," *J. Multimedia*, 2014.

[3] J. Girard, J. Cohn, M. Mahoor, S. Mavadati, Z. Hammal, and D. Rosenwald, "Nonverbal social withdrawal in depression: Evidence from manual and automatic analyses," *IVCJ*, pp. 641–647, 2014.

[4] C. Florea, L. Florea, and C. Vertan, "Learning pain from emotion: Transferred hot data representation for pain intensity estimation," in *CV - ECCV 2014 Workshops*, ser. Lecture Notes in Computer Science, 2015, vol. 8927, pp. 778–790.

[5] J. M Michael Cohen, *Perspectives on the Face*. Oxford, UK: Oxford University Press, 2006.

[6] Z. Zeng, M. Pantic, G. Roisman, and T. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE T. PAMI*, vol. 31, no. 1, pp. 39–58, 2009.

[7] P. Ekman and W. Friesen, *Facial Action Coding System: A technique for the measurement of facial movement*. Palo Alto, U.S.: CA: Consulting Psychologists Press, 1978.

[8] P. Ekman, W. Friesen, and J. Hager, *Facial Action Coding System (FACS) Manual*. U.S.: Research Nexus division of Network Information Research Corporation, 2002.

[9] M. Valstar and M. Pantic, "Fully automatic recognition of the temporal phases of facial actions," *IEEE T. SMC, Part B: Cybernetics,*, vol. 42, no. 1, pp. 28–43, 2012.

[10] P. Liu, S. Han, Z. Meng, and Y. Tong, "Facial expression recognition via a boosted deep belief network," in *IEEE CVPR*, 2014, pp. 1805–1812.

[11] D. Mery and K. Bowyer, "Recognition of facial attributes using adaptive sparse representations of random patches," in *CV - ECCV 2014 Workshops*, ser. Lecture Notes in Computer Science. Springer International Publishing, 2015, vol. 8926, pp. 778–792.

[12] A. Williams, "Facial expression of pain: an evolutionary account," *Behav Brain Sci.*, vol. 25, no. 4, pp. 439–488, 2002.

[13] P. Ekman and E. L. Rosenberg, *What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System*. Oxford, UK: Oxford Univ. Press, 2005.

[14] Y.-L. Tian, T. Kanade, and J. Cohn, "Recognizing action units for facial expression analysis," *IEEE T. PAMI*, vol. 23, no. 2, pp. 97–115, 2001.

[15] W. Zheng, X. Zhou, C. Zou, and L. Zhao, "Facial expression recognition using kernel canonical correlation analysis (kcca)," *IEEE T. NN*, vol. 17, no. 1, pp. 233–238, 2006.

[16] G. Guo and C. R. Dyer, "Learning from examples in the small sample case: Face expression recognition," *IEEE T. SMC Part B*, vol. 35, no. 3, pp. 477–488, 2005.

[17] X. Huang, G. Zhao, M. Pietikainen, and W. Zheng, "Dynamic facial expression recognition using boosted component," in *ACIVS*, 2010, vol. 6475, pp. 312–322.

[18] H. Soyel and H. Demirel, "Improved sift matching for pose robust facial expression recognition," in *IEEE FG*, 2011, pp. 585–590.

[19] P. Lucey, J. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression," in *IEEE CVPRW*, 2010, pp. 94–101.

[20] T.Kanade, J. F. Cohn, and Y. Tian, "Comprehensive database for facial expression analysis," in *Proceedings of the Fourth IEEE FG (FG'00)*, 2000, pp. 46–53.

[21] J. Saragih, S. Lucey, and J. Cohn, "Deformable model fitting by regularized landmark mean-shift," *IJCV*, vol. 91(2), pp. 200–215, 2011.

[22] P. Lucey, J. Cohn, K. Prkachin, P. Solomon, and I. Matthews, "Painful data: The UNBC McMaster shoulder pain expression archive database," in *IEEE FG, pages = 57-64,*, 2011.

[23] M. Valstar, T. Martinez, X. Binefa, and M. Pantic, "Facial point detection using boosted regression and graph models," in *CVPR*, 2010, pp. 2729–2736.

[24] X. Zhu and D. Ramanan, "Face detection, pose estimation, and landmark localization in the wild," in *CVPR*, 2012, pp. 2879–2886.

[25] P. Viola and M. Jones, "Robust real-time face detection," *IJCV*, vol. 57, no. 2, pp. 137–154, 2004.

[26] C.-C. Chang and C.-J. Lin., "LIBSVM : a library for support vector machines," *ACM TIST*, vol. 1, 2011.