Süleyman Demirel Üniversitesi
Fen Bilimleri Enstitüsü Dergisi
Cilt 22, Sayı 1, 160-165, 2018

Süleyman Demirel University
Journal of Natural and Applied Sciences
Volume 22, Issue 1, 160-165, 2018

DOI: 10.19113/sdufbed.50007

An Automatic Multilevel Facial Expression Recognition System

Elena BATTINI SÖNMEZ[*1]

[1]Istanbul Bilgi University, Faculty Engineering and Natural Sciences, Department of Computer Engineering, 34060, Istanbul

(Alınış / Received: 27.10.2017, Kabul / Accepted: 10.02.2018, Online Yayınlanma / Published Online: 16.03.2018)

Keywords
Expression recognition,
Affective computing,
Sparse representation based classifier

Abstract: Facial expression is one of the most natural way of human beings to communicate his-her internal feeling, to stress his-her words, to agree or disagree with the interlocutor, to regulate interaction with the environment and nearby people. This paper challenges the classification experiment run by human beings on the ADFES-BIV database, which is a recently introduced collection of videos expressing low, middle, and high intensity emotions. The proposed automatic system uses the Sparse Representation based Classifier and reaches the top performance of 80 % by considering the temporal information intrinsically present in the videos.

Otomatik Çok Seviyeli Yüz İfadesi Tanıma Sistemi

Anahtar Kelimeler
Yüz ifadesi tanıma,
Duygusal bilişim,
Seyrek yaklaşım tabanlı sınıflama

Özet: Yüz ifadesi, insanoğlunun iç duygusunu ifade etmenin, sözlerini vurgulamanın, muhatabın fikrine katılmanın ya da katılmamanın, içinde bulunulan ortamla ve yakında bulunan insanlarla iletişim kurmanın en doğal yollarından biridir. Bu makalede, yakın zamanda tanıtılan ADFES-BIV video veritabanında yer alan farklı yoğunluk düzeylerinde duygular ifade eden yüzler üzerinde insanlar tarafından yürütülmüş bir sınıflandırma deneyine meydan okuyoruz. Önerilen otomatik sistem Seyrek Temsil Temelli Sınıflandırıcıyı kullanır ve videoların doğası gereği içinde barındırdığı zamansal bilgileri dikkate alarak en iyi performansı olan % 80'e ulaşır.

1. Introduction

Humans are social creatures who communicate, predominantly, via body language; since the face is one of the most expressive part of the body, the study of facial emotional recognition is important for social interaction, and for indicating people's intentions and future actions. In 1872, Darwin theory [1] stated that facial expressions display the internal feelings of a person, and, therefore, they have a relevant communicative role. Given their importance in conveying social information about a person and his-her interaction with the environment, facial expressions recognition is an active research field among researchers in psychology [2, 3, 4, 5, 6, 7].

The corresponding study in the computer engineering field is on automatic facial emotions recognition (FER), which has several applications such as monitoring of disorder condition, human-robot communication, marketing, gaming, etc. However, the implementation of an automatic FER is complicated by the presence of disturbance elements, such as low resolution, scale, illumination, rotation, aging, cultural difference, etc.

At present time, some of the existing datasets with expressive faces store only static photographs, others have videos, where the sequence of frames starts from a neutral face and ends into an emotional one, which is, generally, an acted expression of high intensity. Moreover, many of the databases stores expressions from the six basic emotions of anger, fear, happiness, sadness, surprise, and disgust (8), while others allow for the study of compound or non-basic emotions. Among the most widely used databases are the Cohn Kanade (CK) [9] the Extended Cohn Kanade (CK+) [10], the Japanese Female Facial Expression (Jaffe) [11], the MMI dataset [12], and the Facial Expression in the Wild [13] datasets.

Overall, most of the available dataset store expressive faces at high level of intensity, but emotion unfolds over the time, subtle expressions are more common than exaggerated ones, and an automatic FER system must be able to detect and recognize expressive faces

*Corresponding author: elena.sonmez@bilgi.edu.tr

at different level of intensities, i.e. low, middle and high level. That is, the major obstacle for the implementation of a multilevel emotion recognition system is the availability of databases with validated expressions at different degrees of intensities.

In 2016, Wingenbach et al. [7] introduced to the research community the Amsterdam Dynamic Facial Expression Set-Bath Intensity Variations (ADFES-BIV) database, which is a collection of videos played by non-professional actors, and simulating nine emotions, the six basic expressions plus contempt, pride, and embarrassment, at three levels of intensities, low, medium and high. The creators of the database pointed out on the importance of having videos instead of static images, because the presence of this dynamic stimulus, the temporal progression from a neutral face to the expressive one, is claimed to allow for an easy decoding of the expression [2, 14]. Wingenbach et al. validated this database by assessing the performance of humans in recognizing the acted emotions; they focused on accuracy as well as response time.

The main objective of this study is to build an automatic multilevel emotion recognition system, and to compare its performance against the experiment run by [7]. That is, we are wondering if a machine learning based system behaves in a human-like fashion, and we check it by comparing the accuracy of the automatic system against the one of humans, and by detecting if there is the same trend of accuracy, i.e. lowest recognition rate for low intensity emotions, highest performance for high intensity expressions. Other important contributions of this work are to strength the connection between researchers in psychology and computer engineers, and to introduce the ADFES-BIV database in the engineering environment.

Section 2 presents previous studies on FER with multi levels of intensities, Section 3 presents the ADFES-BIV database, Section 4 introduces the Sparse-Representation based Classifier, Section 5 details the experimental setup and results. Conclusions a drawn in Section 6.

2. Literature Review

Most of previous works done on automatic emotions recognition focuses on high intensity expressions. This is probably due to the difficulty of the task, as well as the shortage of databases with labeled middle and low intensity expressions. In the following there is a selection of studies, which considered the different levels of expressions' intensities.

In 2010, Yang et al. [15] used the location of action units to divide every emotional face into patches, and extracted a compositional feature out of every patch. They compared the results of their experiments on subsets of the CK+ database; more in details, they collected the last frames of every video to create their own set of apex data, while the set of intermediate frames of CK+ videos formed the onset data.

In 2010, Wu et al. [16] explored the use of Gabor Motion Energy filters to detect low intensity facial expressions. The authors worked on a selection of pictures from the CK database; both training and test data were divided into onset and apex images, where onset faces, low intensity expressions, were the first six frames of the video, while apex faces, high intensity expressions, were the last six frames.

In 2011, Jia et al. [17] presented a multi-layer sparse representation (MLSR) algorithm for multi-intensity expressions recognition. MLSR is a block-based SRC working on features extracted via Local Binary Patterns. They tested the proposed method on low and middle intensities expressions, which are intermediate frames of the videos of the CK+ database.

In 2013, Jeni et al. [18] proposed a part-based sparse representation for a continuous measurement of facial action units. They worked with the CK+ database and they considered the first six frames of every videos as onset, or low intensity facial expressions.

In 2017, Surace et al. [19] worked on the group emotion recognition challenge (13), with expressions at multiple levels of intensity. Their best performance was reached with deep neural network and Bayesian classifier.

3. The Amsterdam Dynamic Facial Expression Set-Bath Intensity Variations (ADFES-BIV)

The ADFES-BIV database is an extension of the ADFES database, which was first introduced by Van der Schalk et al. [20]. ADFES is acted by 12 Northern European players (5 female, 7 male) and 10 Mediterranean actors (5 female, 5 male) expressing the six basic emotions plus the three complex emotions of contempt, pride, and embarrassment, and the neutral face.

Wingenbach et al. [7] created the ADFES-BIV dataset by editing the 120 videos played by the 12 Northern European actors to add three levels of intensities. That is, out of every selected tape of ADFES, Wingenbach et al. created three new videos, displaying the same emotion at three different degrees of intensity, low, medium and high, for a total of 360 videos. Every tape of ADFES-BIV starts with a neutral expression and ends with the highest expressive frame. The label of the video gives information of the acted emotion as well as its level of intensity, i.e. low, middle and high. In other words, every video starts with a neutral face, and ends,

respectively, with an expressive face at low, middle, or high intensity, as dictated by its label.

Four of the 12 actors of ADFES-BIV acting the Joy emotion at high intensity are shown in Figure 1; the magenta points are the automatically detected face landmarks that will be necessary to make alignment.

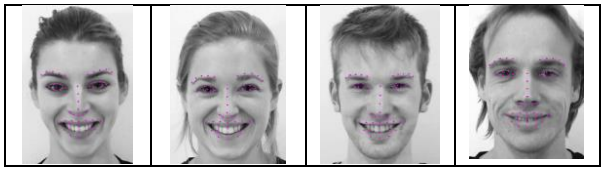

Figure 1. Four of the 12 actors of the ADFES-BIV database expressing Joy emotion at high level.

Figure 1 reveals also the presence of several disturbance elements, which complicate the automatic emotion recognition tasks, such as somatic differences, zoom, pose and illumination.

The first goal of the creators of ADFES-BIV was to validate the newly introduced database by running an emotion recognition experiment, and checking the variation of accuracy rates and response latencies. From the psychological point of view, the database is consistent if low intensity expressions have lower accuracy and higher response latency, when compare to the middle intensity expressions; following the same logic, since high intensity expressions are the easy to recognize, they must have higher hit rate and lower response latency. Wingenbach et al. run all experiments with a sample of 92 participants (51 female, 41 male) recruited from the University of Bath community. The training of the participants was done with 10 extra videos of one Mediterranean actor of the ADFES database. Results of the experiments validated the database: as expected, high intensity expressions were recognized with the top accuracy rates and lowest response latency, whereas expressions with low intensity were recognized with the lowest performance and highest response latency. Since only trials with correct response were used in the calculation of the response time, we did not consider the time variable in our study.

4. The Sparse Representation based Classifier

This work uses the Sparse Representation based Classifier (SRC), which is a successful classification algorithm, first introduced by Wright et al. [21], in 2009. In their study, Wright et al. used SRC for classification of human faces from frontal views with several disturbance elements; SRC proved to be robust against illumination, occlusion and disguise.

SRC codes the input test image as linear combination of all training samples; among all possible solutions, it soughts for the sparsest one. In formula:

$$\hat{x} = armin||x||_1 subject\ to\ D \cdot x = y \qquad (1)$$

where matrix D is a column matrix made up of vectorized training samples, $x \in \Re^M$ is a weights' vector used to create the sparse representation of y, and $y \in \Re^N$ is the vectorized test sample.

Finally, having the sparsest vector $\hat{x}$, the SRC algorithm assigns the test sample y to the nearby class, c, which is the one having minimum distance. In formula:

$$class(y) = argmin_c||y - D \cdot x_c||_2 \qquad (2)$$

A detailed description of SRC is given in [22, 23].

The limited number of available samples makes this database not suitable for Neural Networks, and a comparison study with Support Vector Machine (SVM) is part of our future work.

5. Experimental Setup and Results

The aim of this work is to build and use an automatic facial expression recognition system to challenge the ADFES-BIV database, and to compare the obtained performance against the one of [7]. That is, since computer vision and machine learning algorithms are inspired by human vision and brain, comparing the performance of automatic systems against the one of humans can help to find new paths forward.

In their study, Wingenbach et al. considered both the raw and the unbiased hit rate of success. This is common in the psychological environment, were the recognition task is done by human beings, who can be biased toward some expressions. In other words, the use of unbiased hit rate is necessary to avoid invalid conclusions, in cases of subjects using indiscriminately only one or few response options. Since this practice is not used in the computer engineering field, in the following, we will compare our performance against the raw hit rate of [7].

The average raw hit rate of Wingenbach et al. is 56% for low intensity expressions, 68% for middle magnitude, and 75% for high intensities expressions.

The emotion recognition experiment run by [7] consists in 360 trials presented in a random order to each of the 92 adult participants. At the end of every video, the participant had to select the perceived emotion. Like Wingenbach et al. we worked with 9 emotions plus the neutral face, for a total of 360 videos. Since an automatic system needs to have a training set, we used the Leave-One-Subject-Out (LOSO) technique: having 12 actors, it results in 12 trials; at every trial all videos played by one actor are used for testing, while the remaining videos are the training set. That is, the need of a wider training set forced us to slightly change the experiment presented in [7], but the test set is the same.

Before running the classification experiment, it was necessary to normalize all faces by (1) automatically detect their face landmarks, (2) estimating the coordinates of the left and right pupils, and imposing zero-slop to the line crossing them, (3) imposing a fix inter-ocular distance (IOD) to every face, and (4) cutting and resizing every face to a fix size. The following picture details all steps:

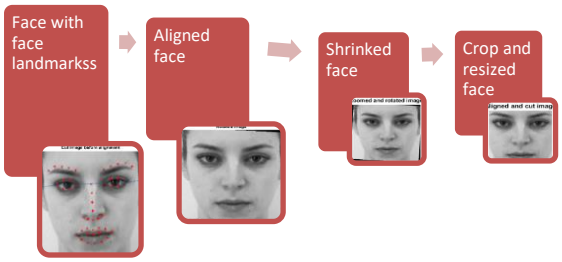

Figure 2. Preprocessing steps.

The first variation of the LOSO experiment considers only the last frame of every video. That is, it does not take advantage of the temporal unfolding of the expression, i.e. the change of the face from neutral to expressive. Results are reported in the first column of Table 1: the average performance is 60% with onset faces, 81% with middle intensities expression and 83% with apex faces. The second variation adds temporal information by subtracting neutral faces from the corresponding peak faces. The second column of Table 1 shows the average performance for low (70%), middle (84%), and high (85%) intensity expressions.

Table 1. Comparison of the accuracy (%) of the two LOSO experiments: without and with temporal information.

Level of intensity	Without temporal information	With temporal information
Low intensity	60	70
Middle intensity	81	84
High intensity	83	85

Results of Table 1 empirically prove that also an automatic emotion classification system is affected by temporal information. That is, the importance of facial dynamics has been investigated and demonstrated by several previous works in psychology [2, 6, 14], and it is, once more, verified in the computer engineering field.

More in details, Table 2 compares the class accuracy reported by [7] against the one obtained by the proposed automatic system. The first column of Table 2 reveals that human beings can easily recognize Surprise and Joy faces, while they find particularly difficult to detect the Contempt and Pride expressions. The accuracy of the implemented automatic system is detailed in the second column of Table 2: Disgust and Angry expressions have the top performance, while Neutral and Contempt are the problematic expressions. Considering the inner difference between the experiment run by Wingenbach et al. [7] and us, we accepted those results.

Table 2. Comparison of the average performance of every class of facial expression.

Performance (%)	Wingenbach et al.	Automatic System
Anger	74	95
Joy	84	86
Disgust	65	97
Fear	62	61
Surprise	92	83
Sadness	79	67
Contempt	34	50
Embarrass	65	89
Pride	42	92
Neutral	89	36

Table 3 details and compares the performance of every class, at every level of intensity: the left section of Table3 presents the raw hit rate of Wingenbach et al., while the right part of Table 3 details the accuracy of the proposed automatic system. Notice that, like [7], we run a 10 classes' experiment working with 9 emotions plus the Neutral faces; that is, our training and test sets contain also the Neutral face; however, like [7], we do not include the Neutral class in the following table because it does not have three levels of intensity.

Table 3. Comparison of the average performance (%) of every class by intensity: (left) the raw hit rate of [7], (right) the performance of the proposed automatic system, (L=Low, M=Middle, H=High).

Emotion	Wingenbach et al.			Automatic System		
	L	M	H	L	M	H
Anger	60	79	85	92	100	92
Joy	68	90	96	58	100	100
Disgust	58	66	71	100	100	92
Fear	51	63	71	58	67	58
Surprise	90	92	95	58	92	100
Sadness	72	82	84	67	67	67
Contempt	27	37	41	42	50	58
Emb.	46	63	85	75	92	100
Pride	30	45	52	83	92	100

The data of Table 3 confirm that the general performance of the automatic system is higher than the one of humans. More in details, the second section of the table, i.e. the accuracy of the automatic system for low, middle and high intensities expressions, shows that Joy, Surprise, Contempt, Embarrass and Pride increase their hit rates together with the level of intensities. On the contrary the performance of Anger, Disgust and Fear drops from middle to high level. This behavior is unexpected, and requires further investigation. Another anomaly of the proposed automatic system is the accuracy of the Sad expression, which is stable to 67%.

5. Conclusion and Future Work

An automatic facial expression recognition system must be able to recognize expressions at different levels of intensity. The recently introduced ADFES-BIV database allows this challenge, because it stores short videos starting with neutral faces and ending with expressive faces, at low, middle and high levels of intensity.

This paper presents an automatic multilevel facial expression recognition system, and challenges the human performance on the ADFES-BIV database. The accuracy of the proposed automatic system is higher than the one of human beings; the use of temporal information extracted from the labelled videos allows the SRC-based system to reach the top accuracy of 80%.

Future work includes (1) to consider the advantages of a block-based approach, that is, to test the performance of the same experimental setup when working with the most discriminative blocks of the face, (2) to investigate on the unexpected results of Table 3, and (3) to make a comparison study with SVM classifier.

References

[1] Darwin, C. 1872. The Expression of the Emotions in Man and Animals. London, England: John Murray; 374 p.

[2] Ambadar, Z., Schooler, J.W., Cohn, J.F. 2005. Deciphering the Enigmatic Face. Psychological Science, 16(2005), 403–410.

[3] Marsh, A.A., Kozak, M.N., Ambady, N. 2007. Accurate Identification of Fear Facial Expressions Predicts Prosocial Behavior. Emotion, 7(2007), 239–251.

[4] Scherer, K.R., Mortillaro, M., Mehu, M. 2013. Understanding the Mechanisms Underlying the Production of Facial Expression of Emotion: A Componential Perspective. Emotion Review 5(2013), 47–53.

[5] Lander, K., Butcher, N. 2015. Independence of Face Identity and Expression Processing: Exploring the Role of Motion. Frontiers in Psychology. 1(2015), 6-255.

[6] Wehrle, T., Kaiser, S., Schmidt, S., Scherer, K.R. 2000. Studying the Dynamics of Emotion Expression Using Synthesized Facial Muscle Movements. Journal of Personality and Social Psychology, 78(2000), 105-119.

[7] Wingenbach, T.S.H., Ashwin, C., Brosnan, M. 2016. Validation of the Amsterdam Dynamic Facial Expression Set – Bath Intensity Variations (ADFES-BIV): A Set of Videos Expressing Low, Intermediate, and High Intensity Emotions. PLoS ONE, 11(2016), e0147112.

[8] Ekman, P. 1992. An Argument for Basic Emotions. Cognition and Emotion. 6(1992), 169–200.

[9] Kanade, T., Cohn, J.F., Tian, Y. 2000. Comprehensive Database for Facial Expression Analysis. 4th IEEE International Conference on Automatic Face and Gesture Recognition (FG), 28-30 March, Grenoble, France, 46–53.

[10] Lucey, P., Cohn, J.F., Kanade, T., Saragih, J., Ambadar, Z., Matthews, I. 2010. The Extended Cohn-Kanade Dataset (CK+): A Complete Dataset for Action Unit and Emotion-Specified Expression. IEEE workshop on CVPR for Human Communicative Behavior Analysis, 13-18 June, San Francisco, CA, USA. DOI: 10.1109/CVPRW.2010.5543262.

[11] Lyons, M., Akamatsu, S., Kamachi, M., Gyoba, J. 1998. Coding Facial Expressions with Gabor Wavelets. IEEE Int. Conf. on Automatic Face and Gesture Recognition, 14-16 April, Nara, Japan, 200–205.

[12] Pantic, M., Valstar, M., Rademaker, R., Maat, L. 2005. Web-Based Database for Facial Expression Analysis. IEEE Int. Conf. on Multimedia and Expo, 6 July, Amsterdam, Netherlands.

[13] Dhall, A. Goecke, R., Joshi, J., Hoey, J., Gedeon, T. 2016. EmotiW 2016: Video and Group-Level Emotion Recognition Challenges. ACM ICMI, 12-16 November, Tokyo, Japan.

[14] Bould, E., Morris, N. 2008. Role of Motion Signals in Recognizing Subtle Facial Expressions of Emotion. British Journal of Psychology, 99(2008), 167–189.

[15] Yang, P., Liu, Q., Metaxas, D.N. 2010. Exploring Facial Expressions with Compositional Features. IEEE Int. Conf. on Computer Vision and Pattern Recognition (CVPR), 13-18 June, San Francisco, CA, USA.

[16] Wu, T., Barlett, M.S., Movellan, J.R. 2010. Facial Expression Recognition Using Gabor Motion Energy Filters. IEEE Computer Society Conf. on Computer Vision and Pattern Recognition Workshops (CVPRW), 13-18 June San Francisco, CA, USA.

[17] Jia, Q. Liu, Y. Guo, H., Luo, Z., Wang, Y. 2011. A Sparse Representation Approach for Local Feature Based Expression Recognition. Int. Conf. Multimedia Technology (ICMT), 26-28 July, Hangzhou, China.

[18] Jeni, L.A., Girard, J.M., Cohn, J.F., De la Torre, F. 2013. Continuous AU Intensity Estimation Using Localized, Sparse Facial Feature Space. 10th IEEE Int. Conf. and Workshops on Automatic

Face and Gesture Recognition (FG), 22-26 April, Shanghai, China.

[19] Surace, L., Patacchiola, M., Battini Sönmez, E., Spataro, W., Cangelosi, A. 2017. Emotion Recognition in the Wild using Deep Neural Networks and Bayesian Classifiers. 19th ACM Int. Conf. on Multimodal Interaction (ICMI'17), November 13–17, Glasgow, UK.

[20] Van der Schalk, J., Hawk, S.T., Fischer, A.H., Doosje, B. 2011. Moving Faces, Looking Places: Validation of the Amsterdam Dynamic Facial Expression Set (ADFES). Emotion, 11(2011), 907–920.

[21] Wright, J., Yang, A.Y., Ganesh, A., Sastry, S.S., Ma, Y. 2009. Robust Face Recognition via Sparse Representation. Transactions on Pattern Analysis and Machine Intelligence, 31(2):210–227.

[22] Battini Sönmez, E. 2013. Robust Classification Based on Sparsity. Lambert Academic Publishing, Germany, 99p, ISBN: 978-3-659-40066-7.

[23] Battini Sönmez, E., Albayrak, S. 2013. A Study on the Critical Parameters of the Sparse Representation based Classifier. IET Computer Vision Journal, 7(2013), 500-507.