Multimedia Tools and Applications (2020) 79:13987–14007
https://doi.org/10.1007/s11042-020-08681-4

Recognition of facial expressions based on CNN features

Sonia M. González-Lozoya[1] · Jorge de la Calleja[2] · Luis Pellegrin[1] · Hugo Jair Escalante[3] · Ma. Auxilio Medina[2] · Antonio Benitez-Ruiz[2]

Received: 17 April 2019 / Revised: 20 November 2019 / Accepted: 17 January 2020 /
Published online: 6 February 2020

Abstract
Facial expressions are a natural way to communicate emotional states and intentions. In recent years, automatic facial expression recognition (FER) has been studied due to its practical importance in many human-behavior analysis tasks such as interviews, autonomous-driving, medical treatment, among others. In this paper we propose a method for facial expression recognition based on features extracted with convolutional neural networks (CNN), taking advantage of a pre-trained model in similar tasks. Unlike other approaches, the proposed FER method learns from mixed instances taken from different databases with the goal of improving generalization, a major issue in machine learning. Experimental results show that the FER method is able to recognize the six universal expressions with an accuracy above 92% considering five of the widely used databases. In addition, we have extended our method to deal with micro-expressions recognition (MER). In this regard, we propose three strategies to create a temporal-aggregated feature vector: mean, standard deviation and early fusion. In this case, the best result is 78.80% accuracy. Furthermore, we present a prototype system that implements the two proposed methods for FER and MER as a tool that allows to analyze videos.

Keywords Facial expression recognition · Facial micro-expression recognition · Convolutional neural networks · Machine learning

1 Introduction

Facial Expression Recognition (FER) is one the most convenient ways to incorporate non-verbal information in many human-behavior analysis, because give some clues of emotional states and intentions. For more than three decades, FER has been attractive for computer vision and machine learning researchers due to real-world applications such as animation, market research, medical treatment, autonomous-driving, surveillance, and many other human-interaction systems.

✉ Sonia M. González-Lozoya
sonia.gonzalez68@uabc.edu.mx

Extended author information available on the last page of the article.

One of the main challenges in a FER task consists of dealing with the great variability of data, i.e. facial expressions can be affected by the level of expressiveness, race or personality [12]. In addition, head pose and illumination conditions are presented when recognizing facial expressions. These factors have been tackled using traditional machine learning algorithms that apply hand-crafted or engineered features as those described in [34] and Gabor [23] that use HOG and LBP to obtain models able to recognize facial expressions across different environments [24]. Nevertheless, new models are required in order to address the large variability of data and to learn effective expression-specific representations.

Recently, methods based on deep learning approaches such as those presented in [19, 20, 25, 26, 30] have obtained better performance than methods that use traditional features. In fact, winners of challenges on FER competitions have used solutions based on deep learning [36, 47]. Nevertheless, a major problem when using deep learning is that large amount of data is required for training good models.

In spite of deep learning achievements in FER approaches, two problems have been identified in the existing databases [30]: 1) low number of images, and 2) images taken from highly controlled environments. These problems have motivated to develop FER methods based on gathering images from the Web [26, 27, 30, 40]. However, the users that annotate the gathered images are not experts, so many images are often incorrectly labeled. In addition, it is necessary to use algorithms for cleaning junk images and to remove mislabeled images (e.g., see [30]). Although there are some efforts to build databases recreating in the *wild* settings such as [3, 4, 8], some Web images have low resolution and they do not have facial landmark points that are necessary for preprocessing, the released facial location and landmarks do not capture the faces in all images correctly, making some training and test samples unusable [25, 26].

Due to problems like those described above, the construction of datasets for FER systems is not a trivial task, this is one of the reasons of the reduced number of instances in the actual FER datasets, for example, see [27]. Besides, many of the actual datasets present images without consider both genders [23] or multiple races [10] as Fig. 1 illustrates. Taking into account the aforementioned problems, this study does not look for generalization using one database and testing in another, e.g., see [24, 25] or just including FER databases built in *wild* settings, e.g. [25]; instead, we want to yield a new and greater database that includes diversity by gender, races, cultures or ages in order to take into account different conditions and environments. For our purpose, we exploit the use of CNN features to integrate five publicly available FER databases (described in Section 3.1) and a sixth image dataset gathered from the Web. Unlike other methods, the one implemented into our proposed method by using CNN-based features models the variability among the considered databases by proposing a methodology that integrates datasets.

Generally, traditional machine learning algorithms applied to handcrafted features provide solutions that do not have the flexibility to generalize a task correctly [24]. Therefore, we have chosen to explore the use of CNN features for our approach, this is a reasonable starting point due to the fact that the best solutions for recent FER-related challenges have been obtained with CNN-based models [36, 47]. Additionally, in order to train a general FER classifier that consider different users, the training dataset should include a large variety of facial expression images, which is not the case of many FER datasets, as is described in [10, 23, 27] there are no more than 60 subjects. Therefore, the models used in our method are trained using a mix of datasets that integrate a greater and diverse dataset.

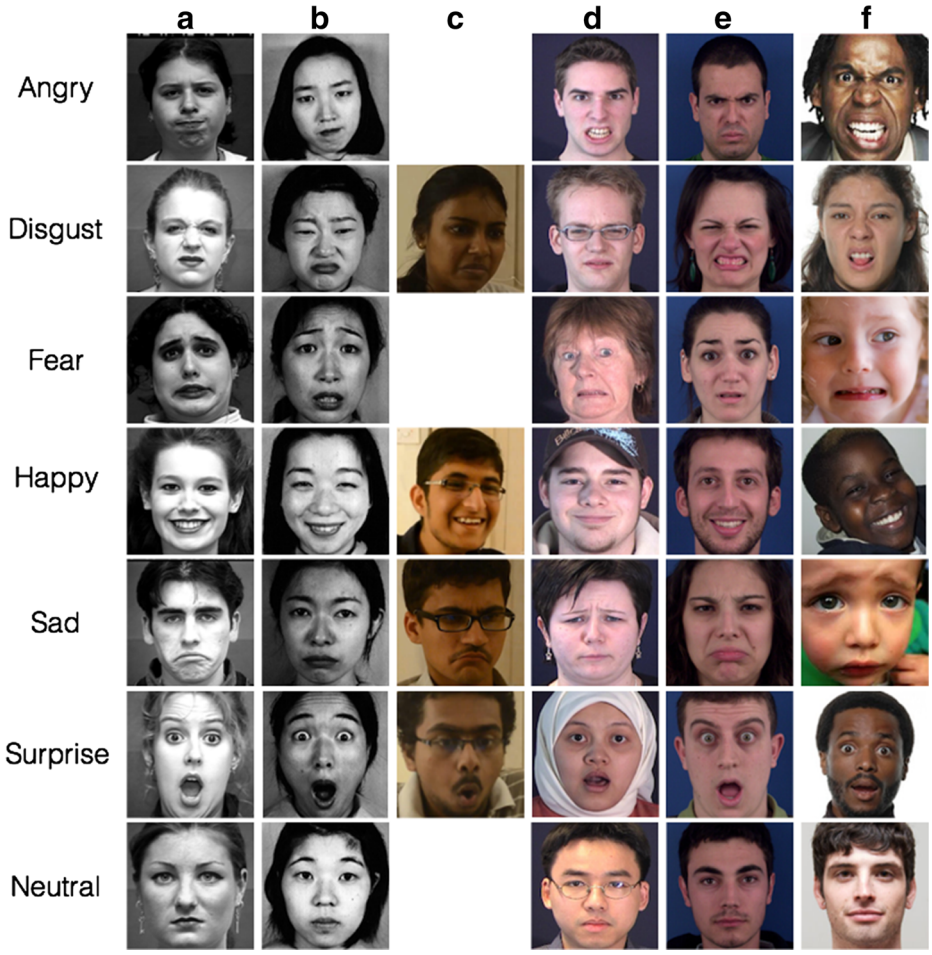


Fig. 1 Sample of images from **a** CK+, **b** JAFFE, **c** ISED, **d** MMI, **e** MUG and **f** Web

The main contributions of this paper are as follows:

- We propose a method to recognize the six facial expressions based on features extracted with convolutional neural networks.
- We show that mixing images from several databases helps to improve generalization.
- We introduce a method to recognize micro-expressions, and we propose three strategies to create a temporal-aggregated feature vector.
- Finally, we present a prototype system that permits to recognize expressions and micro-expressions in videos.

The remainder of the paper is organized as follows. Section 2 reviews the state-of-the-art methods for FER. Section 3 describes the datasets we used in our approach. Section 4 presents the proposed methodology to recognize facial expressions. Experimental results are discussed in Section 5. Section 6 introduces our methodology applied to micro-expressions.

Section 7 presents a prototype system for FE and ME recognition. Finally, Section 8 includes the conclusions.

2 Related work

Some of the FER methods that have been proposed emerged jointly with the introduction of their databases. JAFFE [23] was one of the first FER databases. It was made in a controlled environment, nonetheless is one of the databases less used due to its low variability in images with respect to their visual content. JAFFE authors proposed a FER method based on a set of Gabor filters aligned with the face. MUG [1] is another database made in a controlled environment with little use, its authors presented a FER method [2] that exploits landmark points through manifold representations. Contrary to CK+ [22] database that since its creation has maintaining a constant popularity inside of the scientific community (see Table 2).

In recent years, MMI [27], a Web-based database has been widely used by the community. MMI contains images that were taken in non-controlled environments which attracted the attention of researchers, providing new developments. Numerous works have been dedicated for these two last databases [14, 19, 20, 24, 25, 33, 42, 46, 48]. The study presented in [24] raised the generalization problem faced when one database is used for training and another for testing. Similarly, other researches have taken into account the variability among features from one database to another. For instance, in [19] is proposed a deep architecture specialized in processing of action units (AU), that is independent of the used database. Later, same authors advance their research by means of a dynamic modelling of facial expressions, again using a neural network [20]. A similar work aiming to reduce variability is presented in [42], where a network is designed for different pose faces based on special landmark detection. The work in [46] proposes a weighted mixture deep neural network (WMDNN) that combines features extracted by a pretrained model and classic local binary patterns (LBP). And, a different FER method in [25] follows a same direction at posing that for modelling variability among databases it should to be used a neural network. The proposed method in [33] generates histogram of Gabor filters for representing FE images. Using different approaches, the works in [14, 35] explore the use of sparse representations aiming intra-class variation reduction and preserving structure. In the first work, the FER method relies on the idea that the appearance of intra-class variation for each image could be close to the appearance of the query face image in identity and illumination. While the second one, a dictionary is introduced by decomposing expressions in terms of AUs, where these are coding by sparse representations. Also, looking for person-independent FER, in [48] uses l_p norm via multiple kernel learning for the same end.

A more recent FER database is ISED [10] that consists of images with Indian spontaneous expressions. FER methods introduced in [10] explore traditional features extracted from images with the aim to provide reference results.

Summarizing this section, at the beginning FER methods were dedicated to explore classic features for representing images and exploiting AUs, e.g. see [2, 18, 19, 23], then they identify the need of dealing with variability of subjects. Moreover, the nonexistent generalization among databases brought to create and to explore novel data. In this last path is directed our research. Our aim is to show the usefulness of pre-trained models for modelling data and looking for generalization. The proposed method is simple compared with other in the state-of-the-art but effective, in fact, we show that our methodology could be extended to micro-expression facial recognition (see Section 6).

Next section describe more details about the images in the existing FER databases.

3 Datasets

This section presents a description of the datasets used for experimentation.

3.1 FER datasets

The proposed FER system employs five widely datasets and a sixth set of images from the Web, Fig. 1 shows some examples.

A. The Cohn-Kanade dataset (CK+) [22]. This has 593 sequences of gray-scale images of 123 subjects, but only 309 sequences are labeled. Every image sequence starts showing a neutral face and ends with a facial expression (FE). The size of each image is 640×490 pixels.

B. The Japanese Female Facial Expression (JAFFE) dataset [23]. This consists of 213 images of 60 Japanese females. The size of the gray-scale images is 256×256 pixels.

C. Indian Spontaneous Expression Database (ISED) [10]. This database uses 428 videos obtained from 50 subjects that watch emotional videos. A total of 425 images were extracted. Unlike other datasets, this one only contains four FEs (disgust, happy, sad and surprise). The size of images is 1,920×1,028, thus they have a great resolution.

D. The MMI database [27]. This dataset has 205 image sequences of more than 20 subjects. Each sequence starts and ends with a neutral expression, whereas in the middle of the six FEs is showed. Approximately 500 images of 890×550 pixels were extracted.

E. The MUG facial expression database [1]. This consists of 658 image sequences of 86 subjects. In the same way that MMI database, every image sequence starts and ends with a neutral expression, the FE is in the middle of the sequence; 1,432 images of 896×896 pixels were extracted.

F. Web (our dataset). In order to add new images in the experiments, a three step methodology for gathering images from the Web was implemented. At the first step, Google was used to search images, the keywords used were *face*, *expression*, among others. The images were downloaded in the second step, for the case, we have paid attention to gather images from public domain. Finally, at the third step, an exhaustive manual cleaning step was performed by three experts. The criteria used for removing images were the following: a) consider only images greater than 224×224 pixels; b) check that the images do not belong to the described datasets; c) choose images of relative good quality, that is, with enough illumination, showing a clear face; d) the agreement of the three experts to choose specific images. One hundred images were gathered of each FE approximately, unlike the other databases, the collected image dataset considers different ages and races which allows us to include more diversity than the compiled databases.

In addition to the six traditional facial expressions, we have include images with neutral expressions that were collected from several databases. As we can see in Fig. 1, the considered databases complement each other, i.e., in number of instances, races, cultures, genre, etc. Table 1 shows the number of images (instances) by database. More than 4,000 instances were used in our experiments, which is a considerable number of images and only is overcame by approaches that use thousands of images gathered from the Web [26].

Table 1 Number of images per each expression in databases

	CK+	JAFFE	ISED	MMI	MUG	Web	Total
Angry	45	30	0	151	242	104	572
Disgust	59	29	78	164	242	101	673
Fear	25	32	0	142	183	99	481
Happy	69	31	226	164	259	103	852
Sad	28	30	48	154	202	103	565
Surprise	83	30	73	181	244	106	717
Neutral	123	30	0	0	60	100	313
Total	432	212	425	956	1,432	716	4,173

4 The proposed FER method

The facial expression recognition method consists of three main steps: face detection, CNN-feature extraction and modeling (see Fig. 2). In general, the method works as follows: it takes as input n images from different datasets (described in Section 3). Next, the face contained in the image is found and cropped. After that, these face images are the input to the convolutional neural network to extract relevant features. Finally, the recognition task is through of the modeling performed by a machine learning algorithm. The following subsections describe in detail these steps.

4.1 Face detection

First step is face detection. For performing this task, we rely on the face detector described in [41] that is an implementation of the Viola-Jones (VJ) [38] algorithm. Currently, VJ algorithm shows a good trade-off performance between accuracy and speed in face detection, see Fig. 3 for some examples using random images. Every image is pre-processed and the output is a vector with three pairs of points that indicate the position of a face. Then, the images are cropped covering the entire face and transformed to gray-scale with the purpose of standardized them for the next step.

4.2 CNN feature extraction

The second step uses the VGG-face descriptor [28], a specialized model trained for face recognition by means of a deep learning approach. Figure 4 shows a simplified version of the VGG-face architecture. This model was built with 2.6 million faces obtained from YouTube and with minimal manual labelling. This model is considered due to the size of

Fig. 2 Elements of the FER method. Each element is described in detail in its corresponding section

Fig. 3 Samples of face detection using the method described in [41]

the dataset and because the tasks that implements are very similar to the proposed in our approach.

In order to make this step automatic, we used the MatConvNet [37] framework that enables the integration of the VGG-face model to the previous steps. First, each image is prepared with a final size of 224×224 pixels, scaling or resizing the image as appropriate. Note that in the previous step, the color was removed from the images, although VGG-model works well with color images, we use only gray-scale. The reason of this change was aiming to reduce the variability in illumination over images from FER databases. It is worth to notice that the transformation from RGB to gray-scale images do not reduce performance as well as that CNN models allow us to represent images through neuron activations, where each neuron works as a filter that matches and highlighting a feature such as edges, contour, saturation or gradient endowing CNN features with a great capacity to learn correspondences [21]. Nonetheless, for a practical use in the CNN, every gray-scaled image is tripled to simulate a RGB image. At the end, we have a vector of size of 4,096 values, which describe the face image.

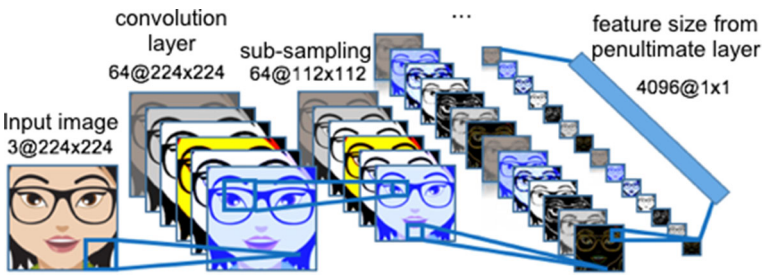


Fig. 4 Simplified version of VGG-face architecture used in our approach

4.3 Modeling

The final step has the aim to learn a model capable to discriminate facial expressions represented by CNN features. For accomplishing this step, we perform a comparison among different traditional machine learning algorithms using Weka [9]. A ten fold-cross-validation experiment was repeated five times changing the seed, then obtaining an average over each algorithm. We found that LibLinear [7] function used in SVM algorithm is superior over the other algorithms, i.e., random forest, classic SVM, KNN, Naïve Bayes and AdaBoost.

5 Experimental results

This section is divided into three subsections, describing the results of the proposed FER method: (1) a comparison vs the state-of-the-art methods, (2) an evaluation of generalization using a *wild* dataset, and (3) results by applying the strategy of mixing databases.

5.1 Comparison vs the state-of-the-art methods

Generally two different scenarios of evaluation are defined: a subject-dependent and a subject-independent. In a subject-independent experiment, databases are split out into training, validation and testing sets in a strict subject independent manner, while subject-dependent scenario partitions do not follow a strict subject-independent. We adopted the subject-dependent scenario for our experiments since CNN representations [13] have the following features: (1) groups of values that often are correlated forming distinctive local motifs that are easily highly detected; and also (2) local statistics of images and other signals are invariant to location.

In Table 2 we show a comparison of the results obtained with our approach, and the ones reported in the literature. In this evaluation, each method is trained and tested using a single dataset. As we can observe, our approach obtains comparable results with those presented in related works; in fact, for the JAFFE database, our method obtains the best accuracy. In average, our approach reaches 92.2% accuracy in the five databases evaluated. However, our aim is to produce a model for predicting new images outside of

Table 2 A comparison of FER accuracy performance (percentages) among the state-of-the-art methods. The best results are in bold

Database	FER methods from the state-of-the-art.
CK+ [22]	94.3 [2]*, 95.8 [11]$^{+}$, 90.4 [14]*, 87.2 [18]*, 92.2 [19]*, 91.4 [20]*, 84.1 [24]*, 93.2 [25]*, 95.7 [33]*, 88.5 [35]*, **98.2** [42]$^{+}$, 97.0 [46]$^{+}$, 95.5 [48]$^{+}$, 89.4 (ours)$^{+}$.
JAFFE [23]	94.7 [14]$^{+}$, 92.2 [46]$^{+}$, **98.26** (ours)$^{+}$.
MMI [27]	93.8 [14]$^{+}$, 74.7 [19]*, 66.2 [20]*, 79.8 [24]*, 77.6 [25]*, 71.8 [33]* 83.1 [42]$^{+}$, 93.6 [48]$^{+}$, **96.4** (ours)$^{+}$.
MUG [1]	**95.8** [2]$^{+}$, 90.6 [2]*, 94.5 (ours)$^{+}$.
ISED [10]	**86.4** [10]$^{+}$, 82.9 (ours)$^{+}$.

$^{+}$Person-dependent scenario. * Person-independent scenario

Fig. 5 Sample of images used for testing generalization performance. Each row corresponds to specific FE and starts with a different color: 'happy' (row 1, yellow), 'disgust' (row 2, black), 'angry' (row 3, red), 'fear' (row 4, purple), and 'sad' (last row, blue)

the databases, therefore we introduce a strategy for mixing databases, which is described below.

5.2 Evaluating generalization

In order to evaluate generalization of the proposed FER method by using an individual database, several experiments were carried out with several models considering databases as follows: individually, jointly and mixed; using different settings.

The methodology to combine two or more databases is the following, taking into account that training is based on binary models: (1) for each FE, the total number of images from the databases to mix are taken as positive instances; (2) while negative instances are defined

Table 3 FER performance accuracy in *wild* dataset by considering different databases, individually or jointly for dataset

	CK+	JAFFE	MMI	MUG
CK+	37.9	43.1	36.3	41
JAFFE	—-	27.2	42.6	42.2
MMI	—-	—-	34.7	34.1
MUG	—-	—-	—-	30.4

Table 4 Number of samples used for different partition training. Set 1 has more negative examples than positive ones, set 2 has a balanced dataset, while set 3 has less negative instances than positive ones

	Set 1		Set 2		Set 3	
FE	$P_{25\%}$	$N_{75\%}$	$P_{50\%}$	$N_{50\%}$	$P_{75\%}$	$N_{25\%}$
Angry	572	1,716	572	572	572	191
Disgust	673	2,019	673	673	673	224
Fear	481	1,443	481	481	481	160
Happy	852	2,556	852	852	852	284
Sad	565	1,695	565	565	565	188
Surprise	717	2,151	717	717	717	239

by the rest of FE images, these are taken randomly from the same databases to combine the same number of positive instances and covering all the remaining FE.

In order to evaluate under the same conditions all settings, we build a new image dataset for testing. These images were gathered from the Web in *wild* conditions (see Fig. 5 for some samples). Please note that this dataset includes images taken under uncontrolled environments, that is, with variations such as illumination, human poses; and also using images with different sizes and illumination. By using a *wild* dataset for testing, defines a more challenging scenario for FER, which was composed by 136 images that were carefully labeled by three experts, including at least 20 images by each FE with a full agreement of labeling.

Table 3 presents the results obtained of this experiment, however, the database ISED was not included due to this only contains four of the six FE evaluated. As we can see, when a unique dataset is used for training, the recognition task considering *wild* conditions results in a complex task. The results for each database that is trained individually are showed in the main diagonal. Here, we can see that CK+ and MMI obtained the best accuracy performance. Instead, the combination of two databases offers better accuracy performance most of the cases. Also, we can see that if more than a database is considered for training, then the limitations on generalization can be alleviated.

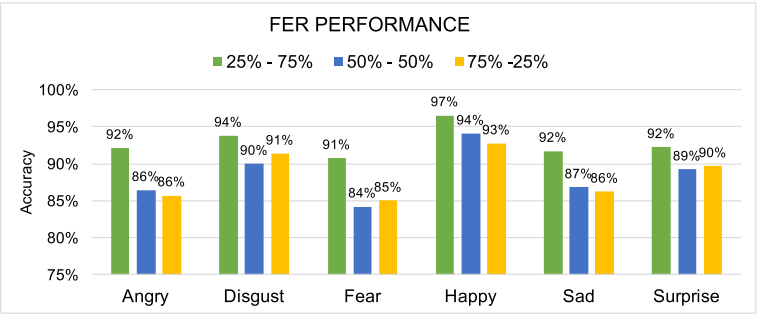


Fig. 6 FER accuracy performance under different partitions for training. The evaluation is carried out using different quantities of positive and negative instances (25%-75%, 50%-50%, 75%-25%)

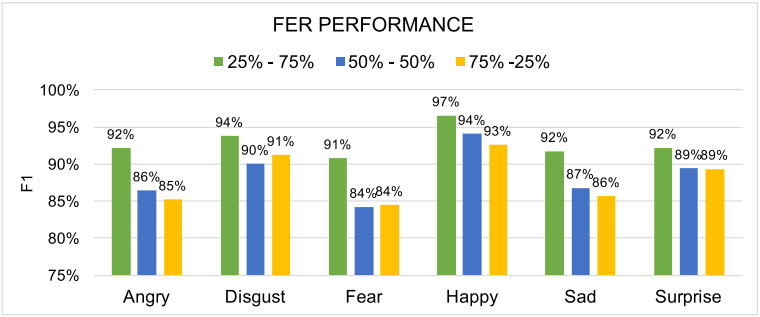


Fig. 7 FER F1 performance under different partitions for training. The evaluation is carried out using different quantities of positive and negative instances (25%-75%, 50%-50%, 75%-25%)

5.3 Evaluating mix databases

First, we experiment with different quantities of instances in order to tune parameters; in our experiment each FE is trained individually, in this sense, a positive example is defined when the FE is presented, and otherwise is negative; these settings are showed in Table 4. As we can see, three settings of the mix databases were tested: (1) 25% of data are positive examples and 75% are negative examples; (2) a half of the data is positive and the other part is negative; and (3) 75% of data are positive instances and 25% are negative instances. Note that the difference between settings rely on the number of negative instances used, while the number of positive instances is fixed. Our aim was to evaluate possible variations when more negative instances are considered.

Figures 6 and 7 show the accuracy and F1 performance results obtained by testing the mixed database with the settings described above. We can observe a consistent behavior between results reported by accuracy and F1 measures. Also, from these results, we can see that the training carried out with the mixed databases using setting 1 (using more negative instances) presents the best performance in all of the six FE evaluated. In particular, we can observe that the *fear* FE is the most challenging recognition task for the settings evaluated.

Table 5 shows the results using the mixed database evaluated in the *wild* dataset. When we compare the results presented before in Table 3, our proposed strategy for mixing databases reaches a better performance, however, it is evident that there is a large room for improving.

Finally, Fig. 8 shows spider graphs that concentrate confusion matrix for each FE. The true positives (TP) are located in the extreme opposite to false negatives (FN), and false positive (FP) in the opposite side of true negative (TN). The idea of the spider graphs is to support the reason of the obtained results in Table 5. In this regard, the results presented

Table 5 FER accuracy performance (percentages) obtained by using the mixed database. S1, S2, and S3 are described in the title of Table 4

	Mix		
Settings	*S1*	*S2*	*S3*
Accuracy	48.8	53.9	45.8

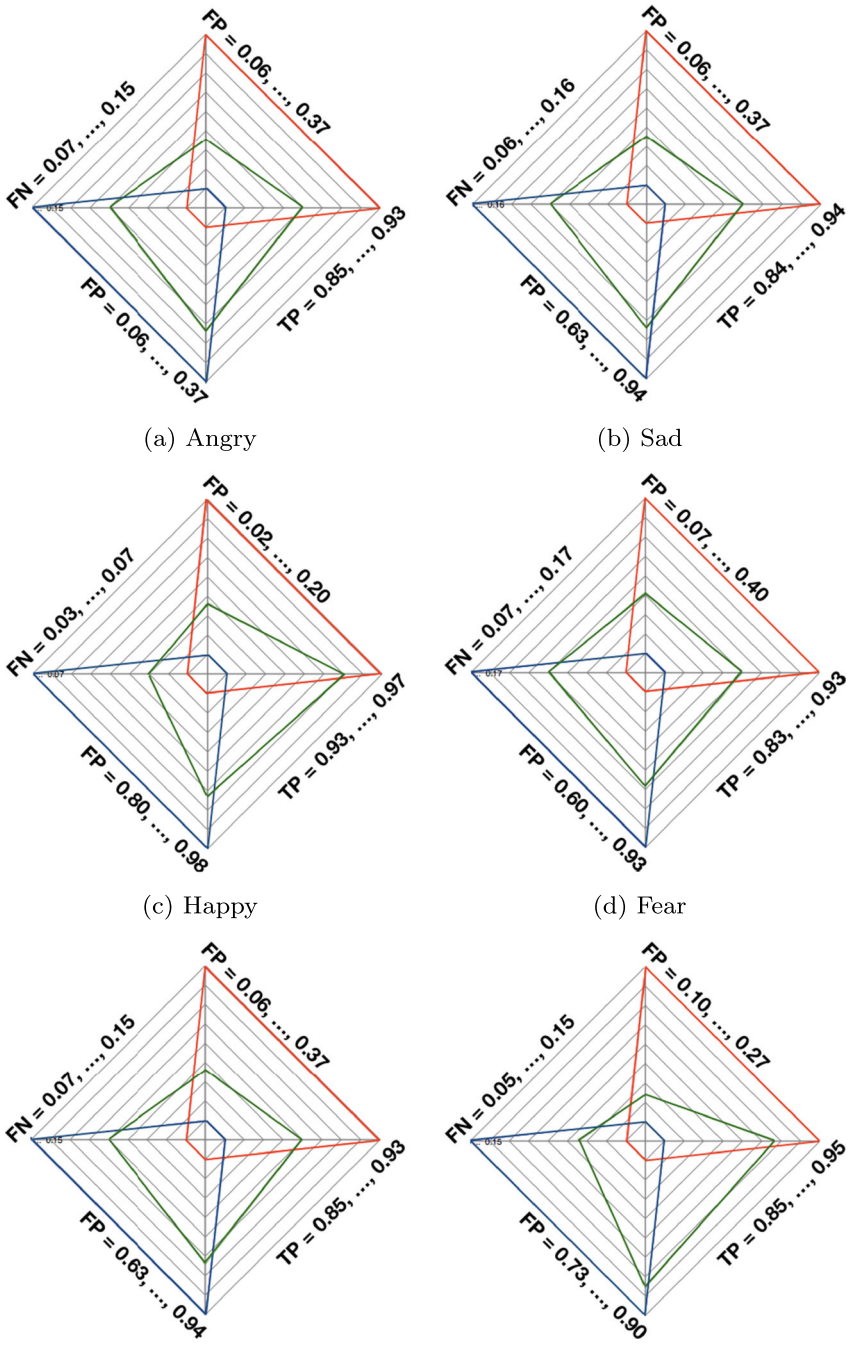


Fig. 8 Spider graphs generated from confusion matrix by each FE under the different training settings considered. Colors in lines indicate different settings: 25%-75% is in blue, 50%-50% is in green, and 75%-25% is red

Table 6 Number of instances used from ME databases

Expression	CASME II [43]	SMIC-HS [17]	TOTAL
Negatives	66	70	136
Positives	34	51	85
Surprise	26	43	69
Non-Micro	0	164	164

by settings 1 and 3 do not show large variations. Instead, the settings 2 shows to be more appropriate for *happy* and *surprise* FEs.

6 Micro expressions for FE recognition

Facial expressions are instantly produced becoming in the most natural way that use people for communicating emotions. A single FE manifestation can directly show effective state, cognitive activity, intentions, personality and psychopathic of a person [5]. When this manifestation is produced unwittingly with a short and low intensity is called a micro-expression (ME), and Ekman mention that these manifestations exposes a person's true emotions [6].

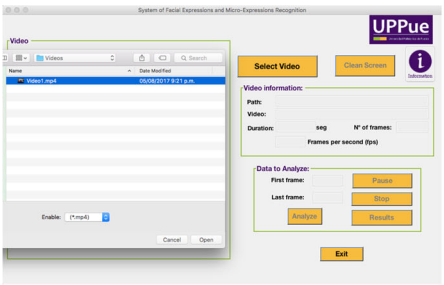
The importance of recognizing ME have motivated to introduce specialized video databases as Polikovsky's dataset [32], SMIC [17, 31], CASME [45] and CASME II [43]. The complexity of each database is based on recognizing ME and classifying from three to five emotions. Different approaches have been proposed [15–17, 29, 31, 39, 43, 44]. Methods as [15, 16, 44] track specific face zones where the ME could be produced, then, model variability aiming to detect and classify the ME.

Motivated by the effectiveness of our proposed method to recognize facial expressions that combines FE images using CNN features, we have experimented to test it for recognizing MEs. The proposed method for ME recognition performs using videos. For this purpose, we have used a combination of databases describe in Table 6. Also, we have tested the method for recognizing micro expressions as negative, positive, surprise and non-micro.

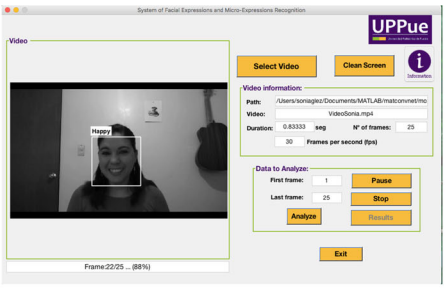
The method we proposed to recognize ME in videos uses three strategies that joints representations of each frame that are extracted from an instance video, into a single one vector. Then, these vectors are used as training set, thence a model is learned for recognizing ME. The three strategies we propose are describe below:

Table 7 Results of accuracy expressed by percentages for recognizing Facial Micro-Expressions using three different strategies

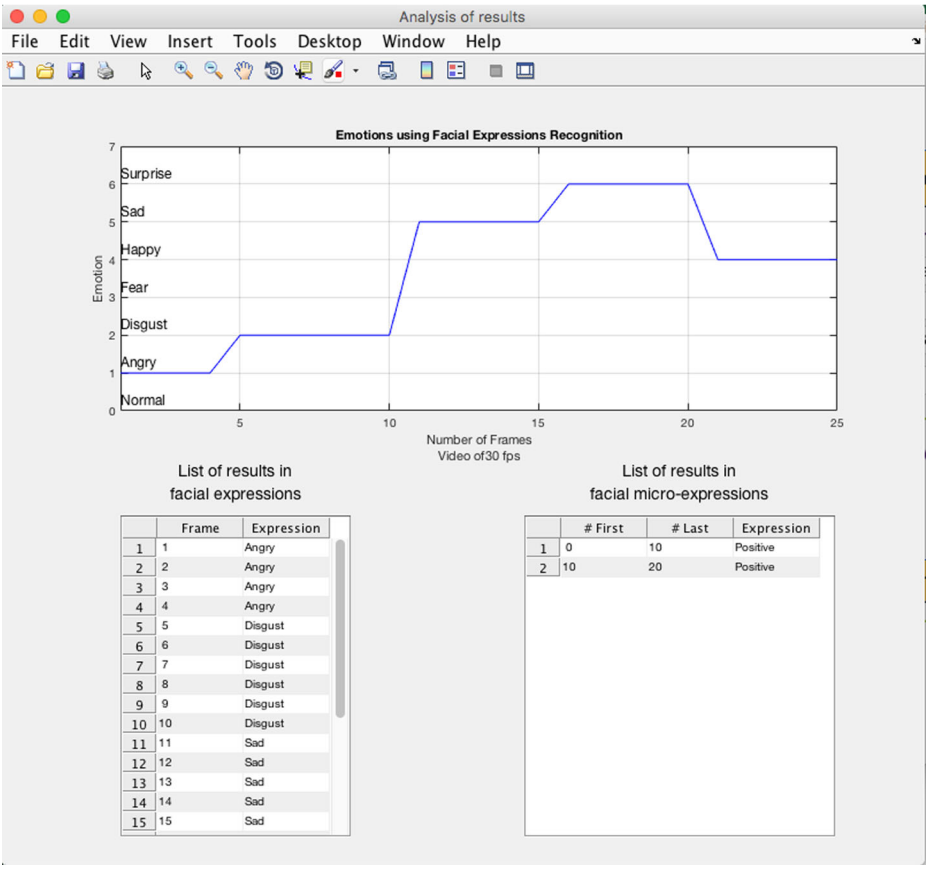
Expression	Mean	Standard deviation	Early fusion
Micro	83.08% ($\pm 1.01^{-2}$)	84.93% ($\pm 1.03^{-2}$)	85.37% ($\pm 6.72^{-3}$)
Negatives	73.45% ($\pm 5.45^{-3}$)	71.10% ($\pm 4.50^{-3}$)	74.62% ($\pm 1.26^{-2}$)
Positives	74.71% ($\pm 1.38^{-2}$)	64.71% ($\pm 2.88^{-2}$)	75.65% ($\pm 1.15^{-2}$)
Surprise	80.72% ($\pm 1.10^{-2}$)	75.65% ($\pm 2.44^{-2}$)	79.57% ($\pm 2.14^{-2}$)
Average	77.99% ($\pm 1.01^{-2}$)	74.10% ($\pm 1.70^{-2}$)	**78.80%** ($\pm 1.30^{-2}$)



(a) Analysis of a video.



(b) Recognition of FE and ME in a video.



(c) Results of the recognition.

Fig. 9 Prototype system developed to recognize FE and ME in videos. **a** A video is selected as input, **b** The video is analyzed frame by frame to recognize facial expressions and micro expressions, the bar below the video shows the progress of processing; **c** The results after analyzing the video are presented in a graphic

1. **Mean.** This strategy consists of averaging features of every frame obtaining a single vector:

$$m \begin{bmatrix} a_{1,1} & a_{1,2} & a_{1,3} & \dots & a_{1,4096} \\ a_{2,1} & a_{2,2} & a_{2,3} & \dots & a_{2,4096} \\ . & & & & \\ . & & & & \\ . & & & & \\ a_{n,1} & a_{n,2} & a_{n,3} & \dots & a_{n,4096} \end{bmatrix} = \begin{bmatrix} m_1 & m_2 & m_3 & \dots & m_{4096} \end{bmatrix} \quad (1)$$

2. **Standard deviation.** In order to analyze the variability among features, we combine each feature of every frame using the standard deviation:

$$std \begin{bmatrix} a_{1,1} & a_{1,2} & a_{1,3} & \dots & a_{1,4096} \\ a_{2,1} & a_{2,2} & a_{2,3} & \dots & a_{2,4096} \\ . & & & & \\ . & & & & \\ . & & & & \\ a_{n,1} & a_{n,2} & a_{n,3} & \dots & a_{n,4096} \end{bmatrix} = \begin{bmatrix} std_1 & std_2 & std_3 & \dots & std_{4096} \end{bmatrix} \quad (2)$$

3. **Early fusion.** Third strategy is based on fusion of two strategies before by using concatenation:

$$\begin{bmatrix} m_1 & m_2 & m_3 & \dots & m_{4096} \end{bmatrix} \cup \begin{bmatrix} std_1 & std_2 & std_3 & \dots & std_{4096} \end{bmatrix} = \begin{bmatrix} m_1 & \dots & m_{4096} & std_{4097} & \dots & std_{8192} \end{bmatrix} \quad (3)$$

Micro expressions recognition was performed using videos, and also we only present the best results, which were obtained using LibLinear function with SVM. Table 7 shows the results, and we can see that standard deviation is better that mean for recognizing micro-expressions, but not for the rest of expressions. An interesting result is that by using these two strategies in an early fusion is beneficial, obtaining the best results of the experiment in both single and average ME recognition.

The results obtained by our MER method could be seen as acceptable, since the results presented by the state-of-the-art methods have an accuracy between 60% to 80% (e.g. see [16, 17, 31, 39, 43, 44]). Hence, we consider our development based on CNN as an alternative to solve such task.

7 Prototype system for FE and ME recognition

We have developed a prototype system that implements the two proposed recognition methods: facial expressions and micro expressions. The prototype takes as input a video, then it is analyzed to find the six universal expressions and micro expressions. The system is able to mark the location where the FE and ME were detected, and outputs the list of all of them. Figure 9 shows the prototype, which can be used for assisting different kinds of evaluations, e.g. in psychological during a session or after of it; providing more information to the specialist, in order to improve decisions about the treatment.

8 Conclusions

We have presented a method for facial expressions divided into three steps: face detection, feature extraction and modeling. Face detection was performed by the well-known

Viola-Jones detector, while a pre-trained convolutional neural network was used for feature extraction. In order to obtain an improvement in generalization for the FER process, we have created a mixed database from several databases. According to experimental results, the best accuracy obtained during training was obtained when using an unbalanced setting composed by 25%/75% of positive and negative instances, respectively. So far, there are no reported results considering a mixed database from several databases. Therefore, the results reported herein are in the state-of-the-art performance. Besides, our FER method goes towards to better FE generalization. Furthermore, to the best of our knowledge, this is the first time that is developed a FER tool focused on reducing the variability of FE. Besides, we have verified that the proposed methodology can be used in ME recognition too, achieving competitive results.

Funding Information This work has been supported by the CONACyT with scholarships No. 71150 and 214764. The authors also would like to thank sponsor *Red temática en Inteligencia Computacional Aplicada* (RedICA).

Compliance with Ethical Standards

Conflict of interests All Authors declare that they have no conflict of interest.

References

1. Aifanti N, Papachristou C, Delopoulos A (2010) The mug facial expression database. In: 11th International Workshop on image Analysis for Multimedia Interactive Services WIAMIS 10, pp 1–4
2. Aifanti N, Delopoulos A (2014) Linear subspaces for facial expression recognition. Image Commun 29(1):177–188
3. Dhall A, Goecke R, Lucey S, Gedeon T (2011) Static facial expression analysis in tough conditions: data, evaluation protocol and benchmark. In: 2011 IEEE International Conference on Computer Vision Workshops (ICCV workshops), pp 2106–2112
4. Dhall A, Goecke R, Joshi J, Wagner M, Gedeon T (2013) Emotion recognition in the wild challenge 2013. In: Proceedings of the 15th ACM on International Conference on Multimodal Interaction, ICMI '13, pp 509–516
5. Donato G, Bartlett MS, Hager JC, Ekman P, Sejnowski TJ (1999) Classifying facial actions. IEEE Trans Pattern Anal Intell 21(10):974–989
6. Ekman P, Friesen WV (1971) Constants across cultures in the face and emotion. J Pers Soc Psychol 17(2):124–129
7. Fan R-E, Chang K-W, Hsieh C-J, Wang X-R, Lin C-J (2008) Liblinear: A library for large linear classification. J Mach Learn Res 9:1871–1874
8. Goodfellow IJ, Erhan D, Carrier PL, Courville A, Mirza M, Hamner B, Cukierski W, Tang Y, Thaler D, Lee D-H, Zhou Y, Ramaiah C, Feng F, Li R, Wang X, Athanasakis D, Shawe-Taylor J, Milakov M, Park J, Ionescu R, Popescu M, Grozea C, Bergstra J, Xie J, Romaszko L, Xu B, Chuang Z, Bengio Y (2015) Challenges in representation learning A report on three machine learning contexts. Neural Netw 64:59–63
9. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH (2009) The weka data mining software: an update. SIGKDD Explorations Newsletter 11(1):10–18
10. Happy SL, Patnaik P, Routray A, Guha R (2017) The indian spontaneous expression database for emotion recognition. IEEE Trans Affect Comput 8(1):131–142
11. Jain S, Hu C, Aggarwal JK (2011) Facial expression recognition with temporal modeling of shapes. In: 2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops), pp 1642–1649
12. Kanade T, Tian Y, Cohn JF (2000) Comprehensive database for facial expression analysis. In: Proceedings of the Fourth IEEE International Conference on Automatic Face and Gesture Recognition 2000, FG '00, pp 46–53
13. Lecun Y, Bengio Y, Hinton G (2015) Deep learning. Nature 521(7553):436–444
14. Lee SH, Plataniotis KN, Ro YM (2014) Intra-class variation reduction using training expression images for sparse representation based facial expression recognition. IEEE Trans Affect Comput 5(3):340–351

15. Li Qiuyu, Zhan Shu, Liangfeng Xu, Congzhong Wu (2019) Facial micro-expression recognition based on the fusion of deep learning and enhanced optical flow. Multimedia Tools and Applications 78(20):29307–29322
16. Li X, Hong X, Moilanen A, Huang X, Pfister T, Zhao G, Pietikäinen M (2018) Towards reading hidden emotions: a comparative study of spontaneous micro-expression spotting and recognition methods. IEEE Trans Affect Comput 9(4):563–577
17. Li X, Pfister T, Huang X, Zhao G, Pietikäinen M (2013) A spontaneous micro-expression database: Inducement, collection and baseline. In: 2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), pp 1–6
18. Littlewort G, Whitehill J, Wu T, Fasel I, Frank M, Movellan J, Bartlett M (March 2011) The computer expression recognition toolbox (cert). In: Face and Gesture 2011, pp 298–305
19. Liu M, Li S, Shan S, Chen X (2013) Au-aware deep networks for facial expression recognition. In: 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), pp 1–6
20. Liu M, Li S, Shan S, Wang R, Chen X (2015) Deeply learning deformable facial action parts model for dynamic expression analysis. In: Cremers D, Reid I, Saito H, Yang M-H (eds) Computer Vision ACCV 2014: 12th Asian Conference on Computer Vision, Singapore, Singapore, November 1-5, 2014, Revised Selected Papers, Part IV. Springer International Publishing, Cham, pp 143–157
21. Long J, Zhang N, Darrell T (2014) Do convnets learn correspondence? In: Proceedings of the 27th International Conference on Neural Information Processing Systems, NIPS'14. MIT Press, Cambridge, pp 1601–1609
22. Lucey P, Cohn JF, 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. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops, pages 94–101. IEEE Xplore
23. Lyons M, Akamatsu S, Kamachi M, Gyoba J (1998) Coding facial expressions with gabor wavelets. In: Proceedings of the 3rd. International Conference on Face & Gesture Recognition, FG '98, pp 200–206
24. Mayer C, Eggers M, Radig B (2014) Cross-database evaluation for facial expression recognition. Pattern Recognition and Image Analysis 24(1):124–132
25. Mollahosseini A, Chan D, Mahoor MH (2016) Going deeper in facial expression recognition using deep neural networks. In: 2016 IEEE Winter Conference on Applications of Computer vision (WACV), IEEE, pp 1–10
26. Mollahosseini A, Hasani B, Salvador MJ, Abdollahi H, Chan D, Mahoor MH (2016) Facial expression recognition from world wild web. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp 58–65
27. Pantic M, Valstar M, Rademaker R, Maat L (2005) Web-based database for facial expression analysis. In: 2005 IEEE International Conference on Multimedia and Expo, pp 317–321
28. Parkhi OM, Vedaldi A, Zisserman A (2015) Deep face recognition. In: Proceedings of the British Machine Vision Conference (BMVC)
29. Peng M, Wang C, Chen T, Liu G, Fu X (2017) Dual temporal scale convolutional neural network for micro-expression recognition. Front Psychol 8(1745):1745–1745
30. Peng X, Xia Z, Li L, Feng X (2016) Towards facial expression recognition in the wild A new database and deep recognition system. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), IEEE, pp 1544–1550
31. Pfister T, Li X, Zhao G, Pietikäinen M (2011) Recognising spontaneous facial micro-expressions. In: 2011 International Conference on Computer Vision, pp 1449–1456
32. Polikovsky S, Kameda Y, Ohta Y (2009) Facial micro-expressions recognition using high speed camera and 3d-gradient descriptor. In: 3rd International Conference on Imaging for Crime Detection and Prevention (ICDP 2009), pp 1–6
33. Sadeghi H, Raie A-A (2019) Human vision inspired feature extraction for facial expression recognition. Multimedia Tools and Applications 78(21):30335–30353
34. Shan Caifeng, Gong Shaogang, McOwan PW (2009) Facial expression recognition based on local binary patterns A comprehensive study. Image Vis Comput 27(6):803–816
35. Taheri S, Qiu Q, Chellappa R (2014) Structure-preserving sparse decomposition for facial expression analysis. IEEE Trans Image Process 23(8):3590–3603
36. Tang Y (2013) Deep learning using support vector machines. CoRR, arXiv:1306.0239
37. Vedaldi A, Lenc K (2015) Matconvnet – convolutional neural networks for matlab. In: Proceedings of the 25th Annual ACM International Conference on Multimedia
38. Viola P, Jones MJ (2004) Robust real-time face detection. Int J Comput Vis 57(2):137–154

39. Wang S-J, Chen H-L, Yan W-J, Chen Y-H, Fu X (2014) Face recognition and micro-expression recognition based on discriminant tensor subspace analysis plus extreme learning machine. Neural Process Lett 39(1):25–43
40. Wang X, Feng X, Peng J (2011) A novel facial expression database construction method based on web images. In: Proceedings of the Third International Conference on Internet Multimedia Computing and Service, ICIMCS '11, pp 124–127
41. Wolf L, Hassner T, Maoz I (2011) Face recognition in unconstrained videos with matched background similarity. In: CVPR 2011, pp 529–534
42. Wu W, Yin Y, Wang Y, Wang X, Xu D (2018) Facial expression recognition for different pose faces based on special landmark detection. In: 2018 24Th International Conference on Pattern Recognition (ICPR), pp 1524–1529
43. Yan W-J, Li X, Wang S-J, Zhao G, Liu Y-J, Chen Y-H, Fu X (2014) Casme ii: an improved spontaneous micro-expression database and the baseline evaluation. PLOS One 9(1):1–8
44. Yan W-J, Wang S-J, Liu Y-J, Wu Q, Fu X (2014) For micro-expression recognition: Database and suggestions. Neurocomputing 136:82–87
45. Yan W-J, Wu Q, Liu Y-J, Wang S-J, Fu X (2013) Casme database: A dataset of spontaneous micro-expressions collected from neutralized faces. In: 2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), pp 1–7
46. Yang B, Cao J, Ni R, Zhang Y (2018) Facial expression recognition using weighted mixture deep neural network based on double-channel facial images. IEEE Access 6:4630–4640
47. Yu Z, Zhang C (2015) Image based static facial expression recognition with multiple deep network learning. In: Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, ICMI '15, pp 435–442
48. Zhang X, Mahoor MH, Mohammad Mavadati S (2015) Facial expression recognition using l_p-norm mkl multiclass - svm. Mach Vis Appl 26(4):467–483

Sonia M. González-Lozoya obtained her B.S. from the Instituto Tecnológico Superior de Cajeme in 2006, and her M.Eng. from the Universidad Politécnica de Puebla in 2017. She is currently doing Ph.D. studies at the Universidad Autónoma de Baja California (UABC). Her research interests are in vision computer and machine learning area.

Jorge de la Calleja is a full-time professor at the Informatics Department at the Universidad Politécnica de Puebla, Mexico, since 2008. He obtained a PhD and MSc in computer sciences from the National Institute of Astrophysics, Optics and Electronics (INAOE). Dr. de la Calleja is member of National System of Researchers from the Mexican Council for Science and Technology (CONACYT). His research interest include machine learning, computer vision and data mining with applications in medicine, education, astronomy and business.

Luis Pellegrin is a full-time professor and researcher of Computer Science in the Faculty of Sciences at the Universidad Autónoma de Baja California (UABC). In 2008, he earned a M.Sc. degree in Artificial Intelligence in the Universidad Veracruzana. And in 2017, he received the Ph.D. in Computer Sciences from the Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE). His research interest is in the vision and language area, including image annotation, automatic generation of sentences, neural networks, and machine learning focus in representation, indexing and automatic analysis of multimodal data.

Hugo Jair Escalante is researcher scientist at Instituto Nacional de Astrofisica, Optica y Electronica, INAOE, Mexico. He holds a PhD in Computer Science, for which he received the best PhD thesis on Artificial Intelligence 2010 award (Mexican Society in Artificial Intelligence). He was granted the best paper award of the International Joint Conference on Neural Networks 2010 (IJCNN). He is a director of ChaLearn, a nonprofit dedicated to organizing challenges, since 2011. He has been involved in the organization of several challenges in computer vision and automatic machine learning. He is reviewer at JMLR, PAMI, and has served as coeditor of special issues in IJCV, PAMI, and TAC. He has served as area chair for NIPS 2016-2018, and has been member of the program committee of venues like CVPR, ICCV, ECCV, ICML, NIPS, IJCNN.

María Auxilio Medina Nieto is teacher and researcher at Universidad Politécnica de Puebla. She has a doctoral and master level from Universidad de las Américas Puebla and a bachelor's degree in computer science from Universidad Autónoma de Puebla (BUAP). At present, she is member of National System of Researchers (SNI) from the Mexican National Council for Science and Technology (CONACYT). Doctor Medina has participated in projects related to institutional repositories, semantic web, databases and human-computer interaction.

Antonio Benitez-Ruiz is teacher and researcher at Universidad Politécnica de Puebla. He has a doctoral and master level from Universidad de las Américas Puebla, and a bachelor's degree in computer science from Universidad Autónoma de Puebla (BUAP). At present, he is the director of graduate programs at Universidad Politécnica de Puebla. Doctor Benitez has interests in mobil robotics, manipulators and image processing.

Affiliations

Sonia M. González-Lozoya[1] · Jorge de la Calleja[2] · Luis Pellegrin[1] · Hugo Jair Escalante[3] · Ma. Auxilio Medina[2] · Antonio Benitez-Ruiz[2]

Jorge de la Calleja
jorge.delacalleja@uppuebla.edu.mx

Luis Pellegrin
luis.pellegrin@uabc.edu.mx

Hugo Jair Escalante
hugojair@inaoep.mx

Ma. Auxilio Medina
maria.medina@uppuebla.edu.mx

Antonio Benitez-Ruiz
antonio.benitez@uppuebla.edu.mx

1 Universidad Autónoma de Baja California (UABC), Ensenada, México

2 Universidad Politécnica de Puebla (UPPuebla), Puebla, México

3 Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE), Puebla, México