



A Novel Approach to Cross dataset studies in Facial Expression Recognition

Silvia Ramis¹ · Jose M. Buades¹ · Francisco J. Perales¹ · Cristina Manresa-Yee¹

Received: 6 October 2021 / Revised: 14 January 2022 / Accepted: 10 April 2022 /
Published online: 29 April 2022

© The Author(s) 2022

Abstract

Recognizing facial expressions is a challenging task both for computers and humans. Although recent deep learning-based approaches are achieving high accuracy results in this task, research in this area is mainly focused on improving results using a single dataset for training and testing. This approach lacks generality when applied to new images or when using it in in-the-wild contexts due to diversity in humans (e.g., age, ethnicity) and differences in capture conditions (e.g., lighting or background). The cross-datasets approach can overcome these limitations. In this work we present a method to combine multiple datasets and we conduct an exhaustive evaluation of a proposed system based on a CNN analyzing and comparing performance using single and cross-dataset approaches with other architectures. Results using the proposed system ranged from 31.56% to 61.78% when used in a single-dataset approach with different well-known datasets and improved up to 73.05% when using a cross-dataset approach. Finally, to study the system and humans' performance in facial expressions classification, we compare the results of 253 participants with the system. Results show an 83.53% accuracy for humans and a correlation exists between the results obtained by the participants and the CNN.

Keywords Convolutional neural networks · Facial expression recognition · Cross-datasets · Facial expression labelling

✉ Silvia Ramis
silvia.ramis@uib.es

Jose M. Buades
josemaria.buades@uib.es

Francisco J. Perales
francisco.perales@uib.es

Cristina Manresa-Yee
cristina.manresa@uib.es

¹ Departament de Matemàtiques i Informàtica, Universitat de les Illes Balears, 07122 Palma de Mallorca, Spain

1 Introduction

Facial expression recognition (FER) has gained increasing interest in the last years due to the high demand of applications for automatic human behavior analysis [3, 7, 13] and novel technologies for human-machine communication and multimedia retrieval [32, 50]. A challenge to solve is that a same expression in different individuals can vary according to ethnicity, age or gender [8, 9, 31]. Although Ekman [8] found that some expressions appear differently across cultures, he also identified seven universal emotions: anger, disgust, fear, sadness, happiness, surprise and contempt. Another feature that can affect is the age, which plays an important role in the representation of emotions. For example, elderly people tend to appear sad or angry when they are in their neutral expression due to the loss of face muscle tone caused by aging [31]. Even the gender can affect, since women generally are more expressive than men [9]. In addition to the human-related factors, there are other factors which also affect face expression recognition. On the one hand, an expression of a particular person may appear differently based on lighting, background, or posture. On the other hand, image-related factors such as image quality, color intensity or resolution, depend on the capture process and environment. These different capture conditions can affect the classification accuracy, especially in cross-dataset evaluations. In most of the published literature, FER is simplified by focusing on optimizing results using the same method or combined-methods on a single dataset or on several datasets separately, but with the training and testing sets belonging to the same dataset [16, 17, 28, 41, 52]. Therefore, these approaches lack of generality when applied to new images or when using it in in-the-wild contexts. This problem can be addressed by combining different datasets for the training, but it is difficult to standardize images from different datasets (regarding image format or capturing conditions).

The main aim of this work is to evaluate how merging information from diverse datasets significantly helps in the training task. Therefore, we present a method to combine multiple datasets to build a large-scale dataset and we conduct an exhaustive evaluation of a proposed system based on a CNN to analyze the performance using a single and cross-dataset approach. Finally, we compare the results both with recent architectures and with humans' recognition.

The main contributions presented throughout this work are: (1) we define a protocol to select and work with different datasets and create a homogenized dataset with data augmentation to be used as a source for a single learning step; (2) we present an extensive evaluation of a proposed CNN using four datasets widely employed in the literature (BU-4DFE, CK+, JAFFE, WSEFEP) and two new datasets (FEGA and FE-Test), using both single and cross-datasets approaches; (3) and we compare the performance of the CNN with state-of-the-art models and with humans' perception. The results show that our approach accurately classifies various facial expressions performing better or at the same level as other state-of-the-art methods and shows a correlation with the humans' classification [23].

The work is structured as follows: In the following section, a review of literature is carried out to identify the most relevant works related to the topic. Section 4 informs about the protocol to create new datasets to train a model based on diverse existing datasets and lists the datasets used in the training and testing of the CNN. This Section also details the image pre-processing and data augmentation steps and describes the proposed CNN for FER. Section 5, 6 and 7 present the exhaustive evaluation conducted to evaluate the pre-processing step, the performance, the comparison with humans and discussion the results. Finally, the last Section concludes the work and summarizes the main contributions.

2 Related work

This Section reviews works on: (1) automatic FER, (2) datasets used for this research area, (3) cross-dataset evaluation and (4) works comparing human results with automatic recognition.

Automatic FER is currently a main area of interest across different fields such as computer science, medicine, or psychology. Research in the area has a long tradition in the Human-Computer Interaction (HCI) discipline and more recently in Human-Robot Interaction (HRI). In the last decades, several techniques have been proposed for FER. Sebe et al. [42] used techniques such as Bayesian networks, Support Vector Machines (SVMs) and decision trees to evaluate several promising machine learning algorithms for emotion detection. SVM was also used by Trujillo et al. [48] for facial expression classification. In [38], the authors studied Gauss–Laguerre (GL) wavelets, which have rich frequency extraction capabilities, to extract texture information of various facial expressions. For each input image, the face area was localized first. Then, the features were extracted based on GL filters, and, finally, a KNN classification was used for expression recognition. Siddiqi et al. [44] used Principal Component Analysis (PCA) and Independent Component Analysis (ICA) for global and local feature extraction, and a hierarchical classifier (HMM) to recognize the facial expression. In [37], Gabor feature extraction techniques were employed to extract thousands of facial features. An AdaBoost-based hypothesis was used to select a few hundreds of the numerous extracted features to speed up classification, and these were fed into a 3-layer neural network classifier trained by a back-propagation algorithm.

More recently, deep learning methods have contributed to improve many research areas [1, 2, 43] and FER is not an exception [6, 16, 17, 20, 27, 28, 41, 45, 52]. Burkert et al. [6] proposed a CNN architecture for FER using the CK+ and MMI datasets in both for training and testing. In [17] the authors proposed a model based on single Deep Convolutional Neural Networks (DNNs), which contained convolution layers and deep residual blocks. Khorrami et al. [20] proposed a CNN for FER. They used the CK+ and TFD datasets and introduced an approach to decipher which portions of the face influenced the CNN's predictions. A combination of CNN and a specific image pre-processing step for the task of emotion detection was proposed in [28], and a Hybrid Convolution-Recurrent Neural Network method for FER in images was presented in [16]. Sajjanhar et al. [41] evaluated the Inception and VGG architectures, which are pre-trained for object recognition. They compared the performance with VGG-Face, which is pre-trained for face recognition. In [45], authors developed a real-time FER system on a smartphone using the CK+, SAIT, SAIT2 and Internet datasets. The Internet dataset was created by the authors; they downloaded face images from the Internet and manually labelled them with five facial expressions. In [52], an ensemble of CNNs was presented with probability-based fusion for FER, where the architecture of each CNN was adapted by using the convolutional rectified linear layer as the first layer and multiple hidden maxout layers. Liu et al. [27] proposed a FER model based on a CNN fused the double-regularized linear support vector machine (L2-SVM).

Mining the literature, we find diverse datasets used for FER, such as CK+ [29], MMI [49], AffectNet [34] or JAFFE [30], being CK+ one of the most popular ones. In Table 1, we summarize the accuracy results reported for some of the architectures developed in this last decade that use the CK+ dataset in their evaluation. Most architectures [6, 20, 27, 28, 33, 41, 45] use *k*-cross-validation [5] to obtain the accuracy results reported in Table 1, except for Jain et al. [17] that performed tests using 98% of the data for training and only 2% for testing.

Sajjanhar et al. [41] used the pre-trained model Face-VGG and Liu et al. [27] used the VGG-11 architecture to perform the feature extraction of human facial expressions. Mollahosseini et al. [33] designed a complex architecture using convolutional layers in parallel and combined them to obtain the final result. Papers [20, 28, 45] presented better results using simpler architectures than papers [17, 27, 33, 41]. Although Burkert et al. [6] presented similar results as the ones reported by Song et al. [45], they used a more complex architecture. Lopes et al. [28] obtained 96.76% accuracy, but the authors only tested with 1 subject for each partition of the k-cross-validation set and ran the experiment 10 times to select the best result. Their method also included a pre-processing step, a CNN using the k-cross-validation method and they reported a value of 89.7% accuracy.

Endeavours addressing the cross-dataset evaluation can be found in [33, 54]. Mollahosseini et al. [33] proposed a deep neural network architecture to address the face FER problem across multiple well-known standard face datasets. The authors evaluated the accuracy of the proposed deep neural network architecture in two different experiments: subject-independent and cross-dataset evaluation. They used six datasets in their experiments. The CK+ dataset was one of the datasets that obtained more accurate results in both experiments: 93.2% and 64.2% respectively. In the work by Zavarez et al. [54], the performance influence of fine-tuning with the cross-dataset approach was investigated. In order to perform this study, the VGGFace Deep Convolutional Network model (pre-trained for face recognition) was fine-tuned to recognize facial expressions. The cross-dataset experiments were organized so that one of the datasets was separated as test set and the others as training set, and each experiment was performed multiple times to ensure the robustness of the results. The authors trained with six datasets and tested with five of them which achieved an accuracy of 88.58%, 67.03%, 85.97%, 48.67% and 72.55% using the CK+, MMI, RaFD, JAFFE and KDEF datasets, respectively.

Tables 2 and 3 summarize the accuracy results performed with recent CNN models for both single and cross dataset approaches. Both tables show the architectures and datasets used, and the results obtained with each test dataset. Most works focus on proposing a CNN and using a single-dataset approach (Table 2), and only a few works deepen in a cross-dataset approach (Table 3). In the case of cross-dataset, we find a similar approach to our work, Zavarez et al. [54] mixed six well-known datasets into one training dataset to test with another well-known dataset.

Table 1 Accuracy results of recent models in the literature. These models were trained and tested with the CK+ dataset to classify 6 basic expressions, except for the models presented in [45] and [27] that were trained to classify 5 and 7 facial expressions, respectively

Model	Year	Accuracy with CK+
Liu et al. [27]	2020	91.58%
Jain et al. [17]	2019	93.24%
Sajjanhar et al. [41]	2018	91.37%
Lopes et al. [28]	2017	96.76%
Mollahosseini et al. [33]	2016	93.20%
Burkert et al. [6]	2015	99.60%
Khorrami et al. [20]	2015	95.70%
Song et al. [45]	2014	99.20%

Table 2 Accuracy results of recent models in the literature that use CNN and the single-dataset approach. These models were used to classify 6 basic expressions, except for the models presented in [45] and [16, 27, 52, 54] that were trained to classify 5 and 7 facial expressions, respectively. The single-dataset method consists of training and testing on the same dataset, where one part is reserved for the testing data and the other for the training data

Model	Year	Architecture	Datasets	Single-Dataset Approach	Cross-Val.	Acc. CK+	Acc. JAFFE	Acc. Faces	Acc. BU-3DPE	Acc. MultiPie	Acc. MMI	Acc. DISFA	Acc. FERA	Acc. SFEW	Acc. FER2013	Acc. TFD	Acc. SAIT	Acc. SAIT2	Acc. Internet Images
Liu et al. [27]	2020	VGGNet + SVM	CK+ JAFFE	✓	✓	91.58	94.81	–	–	–	–	–	–	–	–	–	–	–	–
Jain et al. [17]	2019	Proposed CNN	CK+ JAFFE	✓	✗	93.24	95.23	–	–	–	–	–	–	–	–	–	–	–	–
Sajjanhar et al. [41]	2018	Inception-n-v3, VGG19, VGG-Face Hybrid	CK+ JAFFE Faces	✓	✓	76.82	75.81	82.19	–	–	–	–	–	–	–	–	–	–	–
Jain et al. [16]	2018	Hybrid CNN--RNN	JAFFE, MMI	✓	✗	–	94.91	–	–	92.07	–	–	–	–	–	–	–	–	–
Lopes et al. [28]	2017	Proposed CNN	CK+ JAFFE	✓	✓	96.76	53.44	–	72.89	–	–	–	–	–	–	–	–	–	–
Wen et al. [52]	2017	Hybrid CNN +RNN	BU-3DPE FER2013	✓	✓	–	–	–	–	–	–	–	–	–	66.98	–	–	–	–
Zavarez et al. [54]	2017	VGG16	CK+, JAFFE, MMI, RaFD, KDEF, BU3D-FE, AR	✗	✗	–	–	–	–	–	–	–	–	–	–	–	–	–	–
Mollahosseini et al. [33]	2016	Proposed CNN, AlexNet	Face MultiPie, MMI, DISFA	✓	✓	93.20	–	–	–	94.70	77.90	55.00	76.70	47.70	66.40	–	–	–	–
						92.20	–	–	–	94.80	56.00	56.10	77.40	48.60	61.10	–	–	–	–

Table 2 (continued)

Model	Year	Architecture	Datasets	Single-Dataset Approach	Cross-Val.	Acc. CK+	Acc. JAFFE	Acc. Faces	Acc. BU-3DPE	Acc. MultiPie	Acc. FER2013	Acc. SAIT	Acc. SAIT2	Acc. Internet Images
			FERA, SFEW, CK+, FE-R2013											
Burkert et al [6]	2015	Proposed CNN	MMI, CK+	✓	99.60	–	–	–	–	98.63	–	–	–	–
Khorrami et al. [20]	2015	Proposed CNN	CK+, TFD	✓	95.70	–	–	–	–	–	88.60	–	–	–
Song et al. [45]	2014	Proposed CNN	CK+, SAIT, SAIT2, Internet Images	✓	99.20	–	–	–	–	–	–	97.10	95.50	84.50

Table 3 Accuracy results of recent models in the literature that use CNN and the cross-dataset approach. These models were used to classify 6 basic expressions, except for the models presented in [45] and [16, 27, 52, 54] that were trained to classify 5 and 7 facial expressions, respectively. The cross-dataset method consists of training with one dataset and testing on another dataset. **Acc. one vs one:** accuracy obtained by training with one dataset and testing with a different dataset. **Acc. average vs one:** accuracy obtained by training several datasets separately, taking the average and testing on another dataset. **Acc. mixed vs one:** accuracy obtained by mixing several datasets in a one dataset for the training and testing it on another dataset

Model	Year	Architecture	Num of datasets in the training	Training datasets	Test datasets	Cross-Dataset Approach	Acc. one vs one	Acc. average vs one	Acc. mixed vs one	Test with "Images in the wild"
Liu et al. [27]	2020	VGGNet + SVM	1	CK+, JAFFE	CK+, JAFFE	X	-	-	-	-
Jain et al. [17]	2019	Proposed CNN	1	CK+, JAFFE	CK+, JAFFE	X	-	-	-	-
Sajjanhar et al. [41]	2018	Inception-v3, VGG19, VGG-Face	1	CK+, JAFFE, Faces	CK+, JAFFE, Faces	X	-	-	-	-
Jain et al. [16]	2018	Proposed model	1	JAFFE, MMI	JAFFE, MMI	X	-	-	-	-
Lopes et al. [28]	2017	Proposed CNN	1	CK+	JAFFE, BU-3DFE	✓	45.91	-	-	-
Wen et al. [52]	2017	Hybrid CNN+ RNN	1	FER2013	JAFFE, CK+	✓	50.70	-	-	-
Zavarez et al. [54]	2017	VGG16	6	Trained with all (CK+, JAFFE, MMI, RaFD, BU3DFE, ARFace) except the test set	Emotiv2015, CK+, JAFFE, MMI	✓	76.05	34.09	88.58	-
Mollahosseini et al. [33]	2016	Proposed CNN	1	CK+, MMI	RaFD, KDEF, MultiPie, MMi, DISFA, FERA, SFEW, CK+, FER2013	✓	-	-	44.32	-
									67.03	-
									85.97	-
									72.55	-
									45.70	-
									55.60	-
									37.70	-
									39.40	-
									39.80	-
									64.20	-
	2015		1	CK+, MMI	CK+, MMI	X	-	-	34.00	-

Table 3 (continued)

Model	Year	Architecture	Num of datasets in the training	Training datasets	Test datasets	Cross-Dataset Approach	Acc. one vs one	Acc. average vs one	Acc. mixed vs one	Test iwth “Images in the wild”
Burkert et al [6]		Proposed CNN								
Khorrarni et al. [20]	2015	Proposed CNN	1	CK+, TFD	CK+, TFD	×	–	–	–	–
Song et al. [45]	2014	Proposed CNN	1	CK+, SAIT, SAIT2, Internet Images	CK+, SAIT, SAIT2, Internet Images	×	–	–	–	–

Observing Table 3, few works used a cross-dataset approach for their evaluation and merged several databases. None delved into merged data, nor designed a filtering and data pre-processing protocol to merge them in a homogenized way. Moreover, none have evaluated their system with images “in the wild” or compared results with human perception.

Therefore, it is interesting to study the differences between the perception of a machine and humans. Works such as [46, 47] focused in comparing the performance between machine and humans. On the one hand, Eskil and Benli [47] proposed a set of muscle activity-based features for FER and demonstrated the representative power of the 18 proposed features on three classifiers (Naive Bayes, SVM and Adaboost), in addition to present a comparison between the recognition rates of humans (72%) and their algorithm (77.8%, 87% and 89%, respectively). On the other hand, Susskind et al. [46] designed an experiment with 23 participants to obtain the human performance and compare it with a SVM-based system performance recognizing the six basic facial expressions. They obtained an accuracy of 89.2% and 79.2%, respectively.

For all the above-mentioned reasons we present an extensive study using a CNN in a cross-datasets approach in order to recognize facial expressions in the wild. First, we determine a protocol to select and homogenize different datasets which will use the CNN. In this step, it is important to ensure that the images are correctly labelled, especially in web scrapping datasets. Second, we study the impact on the performance of the facial image pre-processing step, since the majority of papers pay little attention to this step [6, 16, 17, 20, 33, 41, 45, 52, 54]. One of the works that performed an exhaustive study in this step was presented by Lopes et al. [28]. Third, following the research line of studies such as [33, 54], we analyse the performance of the proposed CNN respect to other CNNs and analyse to which extent the use of multiple sources in the CNN’s training phase helps during the test phase. Finally, following works such as [46, 47], we compare the performance of our proposal with human’s perception, and study the similarities and differences between both. Unlike the works [46, 47], we design an experiment with 253 participants and compare the results with the outcomes of the CNN.

3 Materials and methods

In this section, first we present a protocol to create new datasets to train a model based on diverse existing datasets. Second, we describe the datasets used in the training and testing of the Convolutional Neural Network (CNN), including the two new datasets: FEQA and Fe-Test. And third, we propose the FER system as two main modules (see Fig. 1) based on the proposed protocol: a pre-processing step and a CNN.

3.1 Protocol

We can use two methods to build a dataset for facial expressions: using actors or labelling images [10]. The former is based on indicating the actor or actress to pose a facial emotional expression. This method implies the use of multiple actors, and it is a very time-consuming process to capture the images. However, these datasets are considered to include reliable data, they are smaller due to the work involved, but there is the same number of samples for each expression. The later method, labelling images, implies obtaining the images via web browsers and experts classify them. This method depends on the subjective classification of the expert; therefore, it can introduce bias [14].

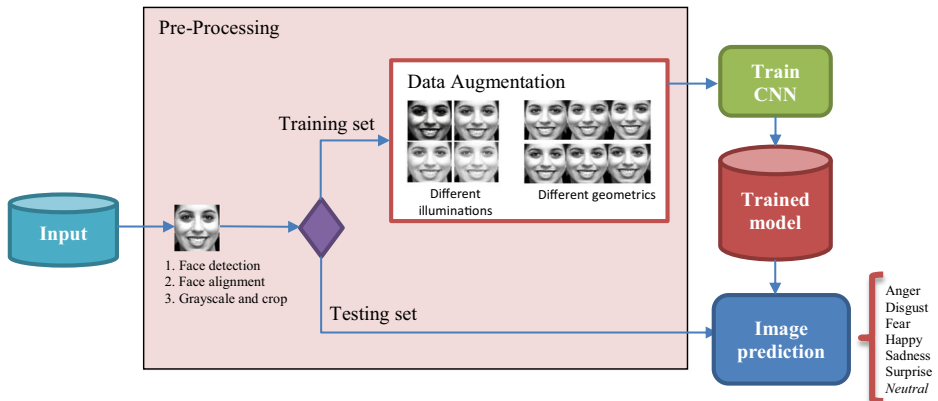


Fig. 1 The FER proposed system scheme. In the first step the input image is processed: face detection and alignment, conversion to grayscale and cropping. If the input image belongs to the test set, the image is processed by the trained CNN and classified in one of the 6 basic expressions, plus the neutral one. If the input image belongs to the training set, a data augmentation is performed on the data and later is processed by the CNN to obtain the trained model

Regardless of how datasets were built, when working with multiple datasets, researchers need to select them, analyze homogeneity criteria across them and assess the image quality. Therefore, we define a protocol to aid the process of datasets selection and data homogenization. We propose a four-step protocol:

Step 1: Dataset filtering. Based on a set of candidate datasets, we need to filter them considering:

- Labelling concordance: the labels must concur across the datasets.
- Similar capture conditions: the capture conditions must be similar or within a range of considered restrictions.
- Lack of duplicity: images must be singular in a dataset and across datasets. The problem of duplicity is prone to occur when datasets are built by web scraping, which can include duplicate or incorrect images.
- Image quality: images must have a minimum resolution and quality.

Step 2: Dataset homogenization. Once the datasets comply with the four requirements, we need to homogenize the images regarding resolution, color space and face alignment. In this way the steps are:

- Face extraction.
- Face alignment and scale: Align and crop the face. Scale images to same resolution.
- Convert image to a common color space.

Step 3: Data augmentation. To augment the data in the dataset, we apply the next techniques found in the literature:

- Gamma correction: Apply different illumination changes.
- Face translation: apply small translations of the face position.

j. Mirroring: as faces are not symmetric, apply mirroring.

Step 4: Train and test the CNN with the homogenized and augmented dataset.

3.2 Datasets

In this subsection, we list the eight datasets used in the study at some stage (see Fig. 2). The first six are standard datasets widely used in facial expression studies: the Extended Cohn-Kanade (CK+) dataset [29], the BU-4DFE dataset [53], the JAFFE dataset [30], the WSEFEP dataset [36], the FER+ dataset [4] and the AffectNet dataset [34]. Then, we built two new datasets: FEGA and FE-test. FEGA is a dataset labelled for the first time, up to our knowledge, with Facial Expression, Gender and Age (FEGA) simultaneously. FE-test, is a new dataset created with images from Internet, which will be used to evaluate FER on images captured “in the wild”.

3.2.1 Datasets built with actors commonly used for facial expression studies

The Extended Cohn-Kanade (CK+) dataset [29] (see Fig. 2(a), first row) contains 593 sequences from 123 subjects ranging from 18 to 30 years old. These sequences were labelled based on the subject’s expression of each of the 7 basic emotion categories: anger, contempt,¹ disgust, fear, happiness, sadness, and surprise.

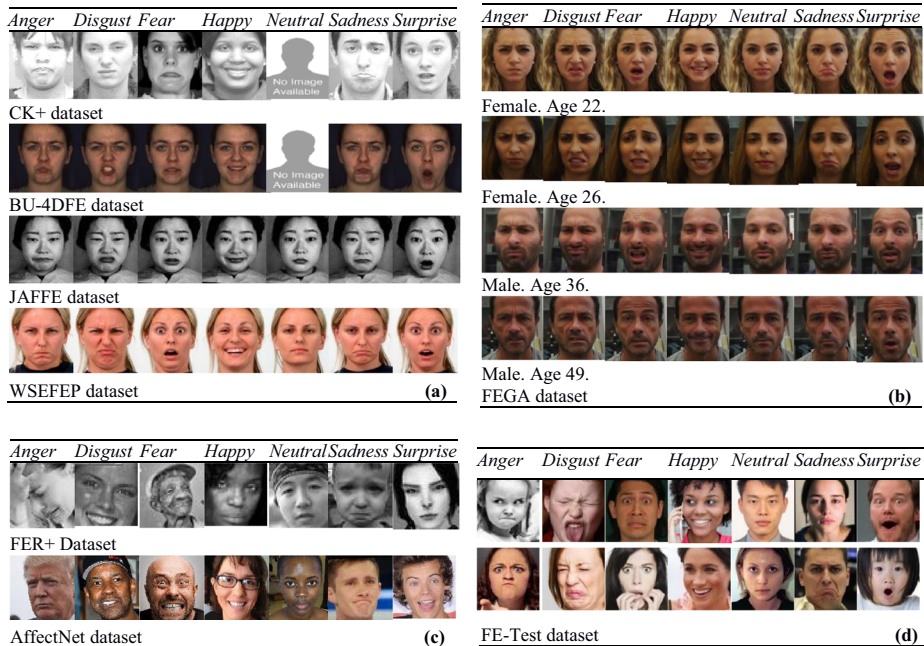


Fig. 2 Different datasets used in this work: (a) Some images of the four popular standard datasets in facial expression (CK+, BU-4DFE, JAFFE and WSEFEP). (b) Excerpt of the FEGA dataset where each row corresponds to a gender and age. (c) Some images of datasets which are built using web scraping. (d) Excerpt of the FE-Test dataset

The BU-4DFE dataset [53] (see Fig. 2(a), second row) contains 606 3D facial expression sequences captured from 101 subjects, 58 females and 43 males. For each subject, there are six sequences showing six prototypic facial expressions (anger, disgust, fear, happiness, sadness and surprise), respectively. The Japanese dataset, JAFFE [30] (see Fig. 2(a), third row), contains 213 images of 7 facial expressions (6 basic facial expressions +1 neutral) posed by 10 female actresses. Each image has been labelled with one out of six expressions. The last dataset is the WSEFEP [36] (see Fig. 2(a), fourth row), which contains 210 high-quality pictures of 30 individuals (14 men and 16 women) posing the basic emotions: happiness, surprise, fear, sadness, anger, disgust and neutral. The images were carefully selected to fit criteria of basic emotions and then evaluated by independent judges.

3.2.2 Datasets built via web scrapping used for facial expression studies

Initially, the FER+ [4] and AffectNet [34] datasets (see Fig. 2(c)) were considered to be included in the study. Both datasets are built using web scraping.

FER+ is an extension of the FER dataset, where the images were re-labelled with 8 facial expressions. This dataset was discarded in the first step of the protocol as it presented a set of limitations. First, the original FER included images with groups of people or even included images that did not correspond to a human face. In addition, classes were not balanced, and each image had only one label. Therefore, they built FER+ by cleaning the dataset removing those images and each image was re-labelled by 10 crowd-sourced taggers, which enables researchers to estimate an emotion probability distribution per face or select the predominant one by majority vote. However, this new version presents limitations too to include it in a cross-dataset evaluation: the image resolution is poor as it is 48×48 . In addition, as the images are obtained by web scraping, there are many duplicates. To analyze this problem, we performed a study of similarities, which was human assisted. First, we carried out an all-to-all comparison using the SSIM algorithm [51]. For each pair of images, we calculated its SSIM value, and when the correspondences between images had a factor greater than 0.7, then the human evaluated them determining if they were the same image. Out of 35,730 images in the database, 30,756 are unique (86.08%) and 4974 are repeated. As an example, an image (id = 1338) appears 31 times (Fig. 3). To analyze the severity of duplicates, we calculated the probability (p_r) of finding at least two repetitions given a sample of N images randomly selected from the dataset. For $N = 32, 64, 128, 256$ and 512 we have $p_r = 0.8786\%, 3.5053\%, 13.4397\%, 43.1406\%$ and 89.0754% respectively. This means that if we train the network with 512 images, approximately 90% (89.0754%) of the time, there will be repetitions. Applying a k-cross-validation algorithm, with $k = 5$, the probability of not having a duplicate in the test and train set is practically zero. Therefore, the problem of duplicates may appear both in the training stage and in the testing stage (with images duplicated in the test and training sets). Due to the poor resolution and the duplicates problem, this dataset was discarded.



Fig. 3 Image repeated 31 times in FER+, whose id are: 1338, 3586, 4501, 5009, 6114, 6122, 9019, 10,242, 10,623, 12,419, 13,190, 13,735, 14,164, 14,240, 15,345, 16,319, 18,779, 19,603, 22,155, 22,226, 25,811, 27,804, 28,660, 29,726, 30,548, 32,880, 33,401, 33,448, 33,622, 33,968 and 34,700

AffectNet is also built using web scraping and images were labelled by 12 experts. To calculate the agreement level between the experts, 36,000 images were annotated by only two human labellers and they agreed in 60.7% of the images. Further, the dataset presents duplicates and a set of images also appear in FER+ dataset. As AffectNet is ten times bigger than FER+, we did not manually analyze the duplicates. However, we studied the exact duplicates applying a hash function (md5) to group the images and looking for exact concordances in each group. Results show that there were 15,244 duplicates, which is 3.63%. Due to the results, this dataset was also discarded.

3.2.3 New datasets (FEGA and FE-test)

This paper introduces the FEGA (Facial Expression, Gender and Age), which was built with 51 subjects, 21 females and 30 males, between 21 and 66 years old. Each subject posed six basic facial expressions [8] (anger, disgust, fear, happiness, sadness, and surprise) plus the neutral face. For each expression and subject, we captured eight RGB images with a resolution of 640×480 pixels. Each subject was labelled with his/her age and gender. The subject performed the seven basic facial expressions, repeating each expression eight times, and one snapshot was taken each time. The images are similar, but not identical, because they were captured at different times. Each image was carefully selected and evaluated by three experts to fit criteria of the seven basic facial expressions and remove the outlier images which did not conform to the required quality for clear perception of expression. Therefore, we present a dataset with 1668 images labelled with facial expression, gender and age, simultaneously (see Fig. 2(b)). We also built FE-Test, which contains 210 frontal images of facial expressions labelled by Google and revised by the research team (see Fig. 2(d)). We chose randomly 30 images from Internet for each expression (anger, disgust, fear, happiness, sadness, surprise and neutral) with different illuminations, backgrounds and image resolution, in addition to faces with different ages and ethnicities. This dataset was labelled by three experts and all images had the agreement of the experts. This dataset was employed to test our algorithm with images “in the wild” obtained from the Internet.

3.3 Image pre-processing and data augmentation

When CNNs are adopted for any task, one of the most neglected steps is the pre-processing one. In fact, the general claim is that a deep model can manage any data variations due to the huge number of parameters (e.g., weights). The basic assumption is the large availability of data: in our case, facial images labelled with the related expression according to the Paul Ekman’s model. However, opposite to the general claim, in this section we highlight the importance of the pre-processing step, which can significantly affect the final network’s performance.

The first step is to detect the face using the method proposed in [26]. Then, we align the images to eliminate rotations and achieve uniformity between them. Eyes positions are found using 68 facial landmarks proposed by [40], which developed the first standardized benchmark for facial landmark localization. To estimate the face’s landmark positions we use the Dlib library, which uses the ensemble of regression trees proposed in work [19]. From these landmarks, we calculate the geometric centroid of each eye and the distance between them. We then draw a straight line (see Fig. 4) to get the angle to rotate the image. The rotation of the axis that crosses both eyes is then compensated to align the eyes horizontally and finally, the face is cropped (see Fig. 4). All images are converted to grayscale within the range 0 to 255 and resized to 150x150pixels.

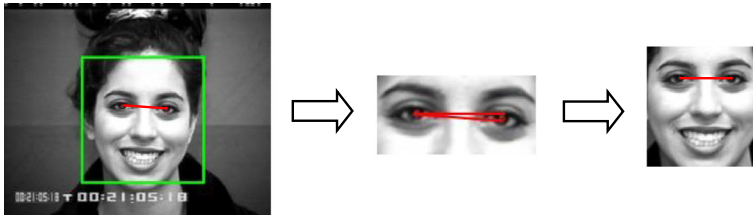


Fig. 4 Left image: face detection and eyes detection. Center image: angle to rotate the image. Right image: face alignment and image cropping

A second important step is to meet the assumption of having a large enough number of training samples that must contain significant facial variations. Initially, the aim of this step confronts with the pre-processing step which seeks to normalize the face images, that is, the variations must be reduced as much as possible. However, it is worth noting that our pre-processing algorithm aims at reducing pose and scale variations, not lighting or appearance ones. Therefore, in the augmentation steps we seek to maintain the basic variations of the input data, and to add new ones in terms of lighting and appearance. With regard to the lighting conditions, we use the gamma correction technique. Eq. (1) is used to adjust the value of gamma,

$$y = \left(\frac{x}{255} \right)^{\frac{1}{\gamma}} \cdot 255 \quad (1)$$

where x is the original image, y is the new image and γ is the value modified to vary the illumination. We use $\gamma = 0.5$, $\gamma = 1.5$ and $\gamma = 2.0$ to obtain a perceptible variation of the original image. For values of γ outside this interval $[0.5, 2.0]$, the face cannot be distinguished. In this way, we quadruplicate the data. Logically, $\gamma = 1$ is not applied, because it does not modify the image (see Fig. 5(a)).

Finally, we introduced some geometric variations, which are aimed at covering for small errors in the position of the eyes during the eyes' location detection. The variations are achieved by translating four pixels in both axes, cropping the image (where the face is always present with two eyes, nose and mouth) and introducing a small appearance variation by duplicating the images through a horizontal flip (see Fig. 5(b)).

3.4 The proposed CNN

Hinton et al. [15] claim that a network with three hidden layers forms a very good generative model of the joint distribution of the images to be classified and their labels. Starting from this base and analysing the different architectures proposed [6, 20, 22, 45], we implement a model with more than three hidden layers to recognize facial expressions. We tune this architecture

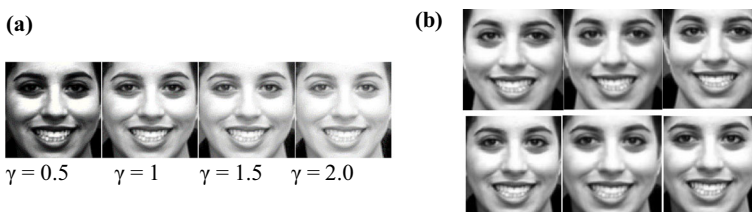


Fig. 5 **a** Images with different illuminations using the gamma correction technique. **b** Images with different geometrics changes. Top row: translation of 4 pixels. Bottom row: the horizontal flip

and its parameters by empirical evidence to improve the results presented in other works [6, 20, 22, 45] (see Table 17).

Table 4 shows the model with three, four, five, and six convolutional layers, using each dataset (BU-4DFE, CK+, JAFFE, FEGA and WSEFEP) separately by means of k-fold cross-validation. It consists of splitting the dataset in k groups and using (k-1) groups as training set and the other one as testing set. We perform k-fold cross-validation using k = 5, since this value have shown empirically to yield test error rate estimates that suffer neither from excessively high bias nor from very high variance [18].

The mean of the accuracies obtained with each dataset after each step of the cross-validation is also shown in Table 4. The highest mean accuracy is obtained with 5 convolutional layers. The addition of more convolutional layers does not improve the results.

The final CNN model is depicted in Fig. 6. Our network receives as input a 150×150 grayscale image and classifies it into one of the next six classes: anger, disgust, fear, happiness, sadness and surprise. The CNN architecture consists of 5 convolutional layers, 3 pooling layers and two fully connected layers. The first layer of the CNN is a convolutional layer that applies a kernel size of 11×11 and generates 32 images of 140×140 pixels. This layer is followed by a pooling layer that uses max-pooling, with a kernel size of 2×2 and stride 2, to reduce the image to half of its size. Subsequently, another two convolutional layers are applied with a kernel of 7×7 and a filter of 32 each. This is followed by another pooling layer, with a kernel size of 2×2 and stride 1, two more convolutional layers that apply a kernel of 5×5 and a filter of 64 each one, and two fully connected layers of 512 neurons each. The first fully connected layer also has a dropout [12] to avoid overfitting in the training. Finally, the network has one output node for each expression that is connected to the previous layer. The output node with the maximum value is classified as the expression of the image.

On the other hand, weight initialization is an important step in neural networks as a careful initialization of the network can speed up the learning process and provide better accuracy results after a fixed number of iterations. Therefore, we carried out a study of the weight

Table 4 Test with different number of convolutional layers to classify 6 classes (one per facial expression). The best result is in bold text with high mean

Number of Convolutional layers	Accuracy (CK+)	Accuracy (BU-4DFE)	Accuracy (WSEFEP)	Accuracy (JAFFE)	Accuracy (FEGA)	Mean
3	90.14	70.57	90.56	64.08	71.01	77.27%
4	90.43	70.28	90.00	64.08	71.24	77.21%
5	91.51	71.49	87.22	63.38	76.06	77.93%
6	91.29	69.17	89.09	62.88	73.98	77.28%

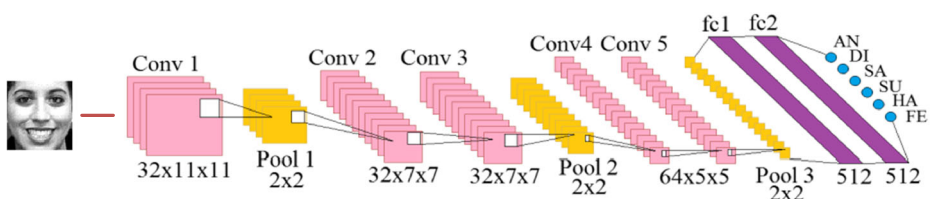


Fig. 6 Architecture of the CNN proposed with 5 convolutional layers, 3 pooling layers and 2 fully connected layers

Table 5 Test with different initializations for each dataset. The mean and standard deviation is shown for each initialization

Initialization	WSEFEP	BU-4DFE	FEGA	JAFFE	CK+	Mean	σ
xavier	87.22	71.49	76.06	63.38	91.51	77.93%	7.11
gaussian	85.00	69.29	71.36	51.23	91.91	73.76%	6.70
msra	90.00	68.89	67.87	60.63	90.47	75.57%	6.72
gaussian+msra	91.11	71.05	75.42	60.62	93.15	78.27%	6.55
gaussian+xavier	91.11	69.61	71.43	54.10	91.44	75.54%	7.23
xavier+gaussian	90.55	72.15	70.63	59.47	92.53	77.07%	7.15
xavier+msra	88.33	69.96	72.52	61.68	90.62	76.62%	7.46
msra+gaussian	88.33	68.45	72.27	57.82	90.82	75.54%	6.03
msra+xavier	86.67	68.84	69.85	59.72	90.95	75.21%	5.82

initialization techniques most used in CNNs. In Table 5 we show accuracy results with different initializations of weights. These initializations consist of combinations of Xavier [12], MSRA [25] and Gaussian [11] methods. The Gaussian method uses a standard deviation of 0.01. We trained our CNN using k-cross-validation (with $k = 5$) with each initialization method. As shown in Table 5, the Xavier method and, the combination of Gaussian and MSRA methods result in higher average accuracy values (marked in bold).

Moreover, the Gaussian+MSRA initialization obtains a lower standard deviation compared with the Xavier initialization, meaning that all the accuracy values are close to the average. For these reasons, we decided to use, in all our experiments, the Gaussian+MSRA initialization (i.e., a Gaussian filler is used for the convolutional layers and a MSRA filler for the fully connected layers). The loss is calculated using a logistic function of the softmax output as in several related works [35, 41, 52]; the activation function of the neurons is a ReLU, which generally learns much faster in deep architectures, and the method used to calculate the weights between neurons is the Adam method [21], since this method shows better convergence than other methods.

3.5 Comparison between our CNN architecture and other recent architectures in the literature

We have compared our architecture with recent proposals in the literature (see Table 6). Note that in [6, 17, 33] the architectures are more complex than the rest. In [17], the authors use 6

Table 6 Results of recent models in the literature. These models have been trained and tested with the CK+ dataset to classify facial expressions. *The authors used a more complex architecture

Model	[17]	[28]	[6]	[33]	[41]	[22]	[20]	[45]	Our Model
Images	128×96	32×32	224×224	48×48	224×224	224×224	96×96	96×96	150×150
LRN ^a	No	No	Yes	No	No	Yes	No	No	No
Convolutional layers.	6+2*	2	1+2*	2+3*	13	5	3	4	5
Pooling layers	3	2	5	4	5	3	3	4	3
Dropout	2	0	1	0	2	2	1	2	1
FC layers	2	1	1	2	3	3	1	1	2

^a Local Response Normalization (LRN) is a layer that square-normalizes the pixel values in a feature map in a local neighborhood

convolutional layers and 2 residual blocks which consist of 4 convolutional layers. The architecture presented in [6] uses 1 convolutional layer and 2 blocks. Each block consists of a parallel path: the first path uses 2 convolutional layers, and the second path uses 1 pooling layer and 1 convolutional layer. In [33], they use 2 convolutional layers and 3 modules which consist of 4 parallel convolutional layers.

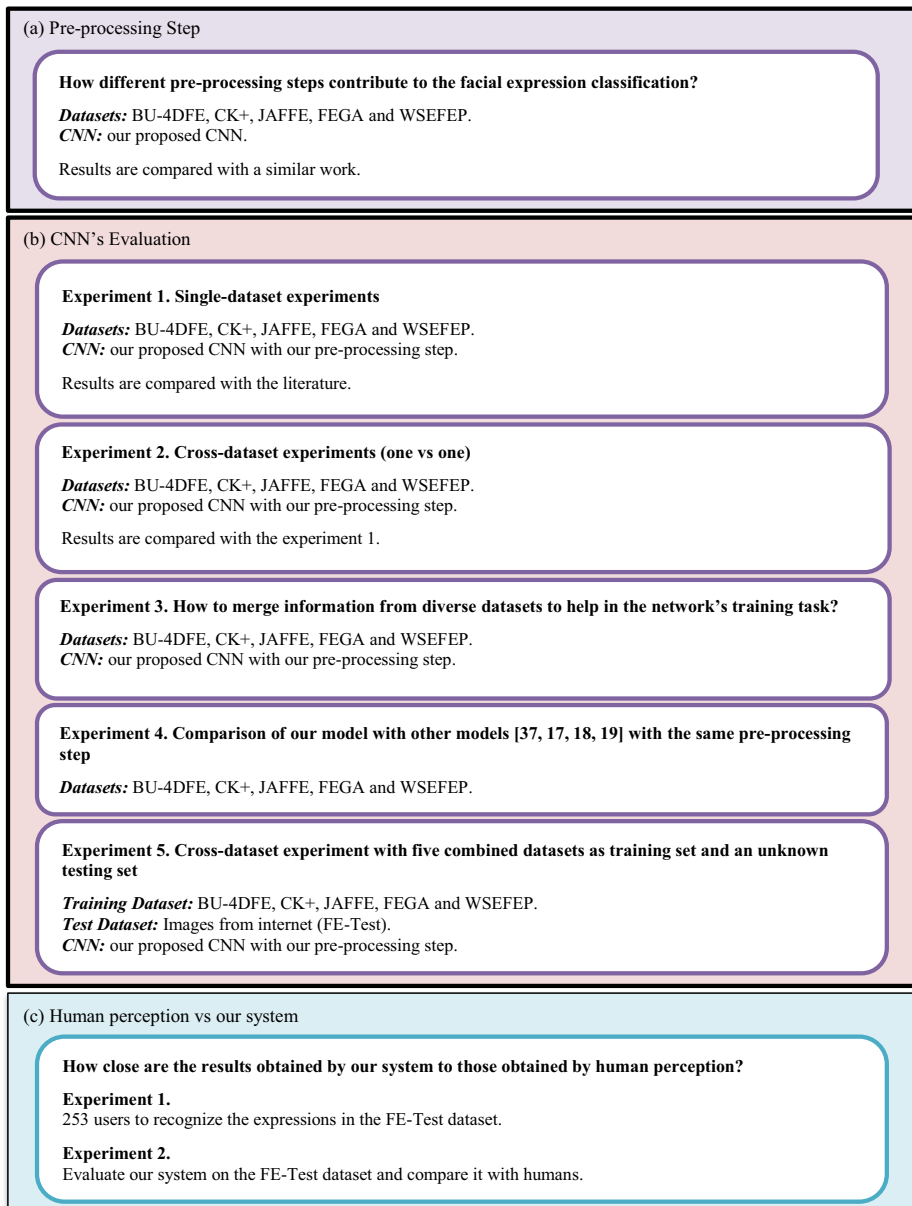


Fig. 7 Summary of the set of experiments performed in this work: (a) pre-processing step evaluation; (b) CNN's evaluation; (c) Evaluation of our system with human perception

Table 7 Test with different image pre-processing with CK+ dataset. Pre-processing steps: (a) no pre-processing, (b) face alignment and crop, (c) face alignment, crop and illumination variations, and (d) face alignment and crop, illumination variations and geometric changes

Pre-processing step	BU-4DFE	FEGA	JAFFE	WSEFEP	CK+	Mean
(a)	50.54	26.50	40.08	35.55	80.15	45.57%
(b)	62.12	63.48	50.65	75.56	82.11	66.78%
(c)	64.48	69.60	51.20	87.22	85.45	71.59%
(d)	72.98	75.42	60.05	88.33	93.15	77.99%

4 The pre-processing step evaluation

In this Section we present an evaluation of the image pre-processing step, where we show that the pre-processing step is relevant to improve the performance despite the intrinsic complexity of a CNN. In this step, we show how different pre-processing steps contribute to the facial expression classification accuracy by adding one at a time. Figure 7(a) summarizes the experiment performed in the pre-processing step identifying the used datasets and CNN.

4.1 Procedure

In this experiment we use our CNN model with each dataset (BU-4DFE, CK+, JAFFE, FEGA and WSEFEP) separately, using 80% of the images as training set and the other 20% as testing set. The training was performed using k-cross-validation with $k = 5$ and only 60 epochs each time. The combinations of pre-processing steps evaluated are: (a) original images without image pre-processing, (b) face alignment and crop, (c) face alignment, crop, and illumination variations using gamma correction technique, (d) face alignment, crop, illumination variations, and geometric changes. Each pre-processing step has been used with all datasets and the six basic expressions (see Table 7). Finally, we compare our results with the results presented in [28], which employs a similar image pre-processing and the CK+ dataset (see Table 8).

4.2 Results and discussion

The results are shown in Tables 7 and 8. Table 7 reports that just the use of the image alignment adds a great improvement to the classification accuracy in all cases. But the best results are achieved with the last option (that incorporates all the pre-processing steps), which improves the results in each step an average of 8.1% with respect on the previous one, having an average improvement of 32.4% between the first step (a) and the last (d).

In Table 8 we compare our results with the results presented in the work [28], which used the CK+ dataset for their experiment. In [28], the used image pre-processing was alignment of the face, crop (only the face without hair), down-sampling of the face image to 32×32 pixels, normalization of the image intensity, and generation of 30 more samples (see Table 8, step (h)). Although both image pre-processing steps are quite similar, the main difference is that we apply the horizontal flip and vary the illumination in order to get more diversity of data instead of applying down-sampling and normalize the image intensity [28]. The results show that our proposed pre-processing step improves an average of 3.45% with respect to those reported in [28].

Table 8 Comparison of results with a similar work in the literature [28] to classify the six basic facial expressions: Anger (AN), Disgust (DI), Fear (FE), Happiness (HA), Sadness (SA) and Surprise (SU). Pre-processing steps: (d) face alignment, crop, illumination variation and geometric changes, and (h) face alignment, crop, down-sampling, normalization and generation of 30 samples more

Pre-processing step	AN	DI	FE	HA	SA	SU	Mean
(d) Model proposed	95.96	97.41	82.69	100	84.88	97.93	93.15%
(h) [28]	79.30	94.40	73.10	99.40	72.80	94.90	89.70%

5 The CNN's evaluation

In this Section we present five experiments to evaluate the accuracy of our system (our proposed CNN with image pre-processing) and we also evaluate the accuracy with the new dataset FEGA. Figure 7(b) summarizes the set of experiments performed in this work with the CNN.

5.1 Experiment 1. Subject-independent evaluation

The aim of this experiment is to present a comparative study between our results classifying the six basic expressions with the recent studies in the literature. The results obtained with each dataset are compared with studies that used the same dataset.

Table 9 (a) Comparison of subject-independent results with related works to classify 6 expressions. The best results are shown in bold text. **(b)** Comparison between experiment 1 and 2. Experiment 1: Subject-independent evaluation. Experiment 2: Cross-datasets evaluation (one dataset as training set against four datasets as testing set). Both experiments classified 6 expressions. The best results are shown in bold text

Training Set	(a) Subject-independent experiment				(b) Cross-datasets experiment	
	Test Set	Accuracy (our Model)	Model	Accuracy (in the literature)	Test Set	Accuracy (our Model)
BU-4FDE	BU-4FDE	73.58%	CNN [28]	72.89%	FEGA	24.96%
					JAFFE	43.17%
					WSEFEP	62.22%
					CK+	47.90%
FEGA (New dataset)	FEGA (New dataset)	72.61%	–	–	BU-4DFE	44.67%
					JAFFE	37.70%
					WSEFEP	73.89%
					CK+	78.19%
JAFFE	JAFFE	56.22%	CNN [17]	95.23%	BU-4DFE	32.49%
			CNN [28]	53.44%	FEGA	23.42%
			CNN [16]	94.91%	WSEFEP	26.67%
			CNN [41]	86.67%	CK+	42.97%
WSEFEP	WSEFEP	87.22%	LBP+KNN [35]	80.00%	BU-4DFE	45.61%
			LBP+SVM [35]	78.10%	FEGA	36.55%
					JAFFE	34.43%
					CK+	59.22%
CK+	CK+	93.15%	CNN [17]	93.24%	BU-4DFE	48.60%
			CNN [28]	96.76%	FEGA	48.45%
			CNN [33]	93.20%	JAFFE	23.50%
			CNN [41]	91.37%	WSEFEP	67.78%

Table 10 Results of classification of the different datasets. We use the five combined datasets (BU-4DFE, FEAGA, JAFFE, WSEFEP and CK+) as training set and we test it with each dataset applying k-fold cross-validation

Train	BU-4DFE	FEAGA	JAFFE	WSEFEP	CK+	Mean
5 dB	100	98.56	98.29	95.55	95.93	97.67%

5.1.1 Procedure

In this experiment, our CNN model is evaluated with each dataset (BU-4DFE, CK+, JAFFE, FEAGA and WSEFEP) separately by means of k-fold cross-validation using $k = 5$ and 60 epochs each time. The pre-processing step (explained in subsection 3.3) is the same for all datasets. Finally, we compare our results with related works to classify 6 expressions (see Table 9(a)) in terms of accuracy.

5.1.2 Results and discussion

The results are shown in the experiment (a) of Table 9. Although the results with the JAFFE dataset can be improved, we demonstrate empirically that our results are competitive compared with other recent and successful works. The JAFFE dataset contains only 10 actresses and needs more data augmentation to achieve good results. However, as shown in experiment (a) of Table 9, we get better results for the BU-4DFE and WSEFEP datasets and for the CK+ dataset, our results are close to the ones published in [17, 33]. Evidently, the FEAGA dataset

Table 11 Results with different combinations of datasets for the testing set of BU-4DFE (left) and FEAGA (right)

Training Set	Test Set	Accuracy	Training Set	Test Set	Accuracy
Without combining datasets					
FEAGA	BU-4DFE	44.67%	BU-4DFE	FEAGA	24.96%
JAFFE		32.49%	JAFFE		23.42%
WSEFEP		45.61%	WSEFEP		36.55%
CK+		48.60%	CK+		48.45%
2 Combined datasets as training set					
WSEFEP & CK+	BU-4DFE	54.36%	WSEFEP & CK+	FEAGA	54.79%
FEAGA & CK+		52.04%	BU-4DFE & CK+		51.47%
FEAGA & WSEFEP		52.19%	BU-4DFE & WSEFEP		39.88%
JAFFE & CK+		52.82%	JAFFE & CK+		48.61%
FEAGA & JAFFE		54.21%	BU-4DFE & JAFFE		26.66%
WSEFEP & JAFFE		43.55%	WSEFEP & JAFFE		38.56%
3 Combined datasets as training set					
WSEFEP, CK+ & JAFFE	BU-4DFE	57.72%	WSEFEP, JAFFE & CK+	FEAGA	48.92%
FEAGA, WSEFEP & CK+		53.61%	WSEFEP, CK+ & BU-4DFE		51.16%
FEAGA, JAFFE & WSEFEP		50.32%	BU-4DFE, JAFFE & WSEFEP		42.43%
JAFFE, FEAGA & CK+		54.06%	JAFFE, BU-4DFE & CK+		46.06%
4 Combined datasets as training set					
WSEFEP, CK+, JAFFE & FEAGA	BU-4DFE	57.68%	WSEFEP, CK+, BU-4DFE & JAFFE	FEAGA	55.02%

Table 12 Results with different combinations of datasets for the testing set of JAFFE (left) and WSEFEP (right)

Training Set	Test Set	Accuracy	Training Set	Test Set	Accuracy
Without combining datasets					
FEGA	JAFFE	37.70%	BU-4DFE	WSEFEP	62.22%
BU-4DFE		43.17%	JAFFE		26.67%
WSEFEP		34.43%	FEGA		73.89%
CK+		23.5%	CK+		67.78%
2 Combined datasets as training set					
WSEFEP & CK+	JAFFE	46.45%	CK+ & FEGA	WSEFEP	82.22%
FEGA & CK+		41.53%	BU-4DFE & CK+		78.33%
FEGA & WSEFEP		26.23%	BU-4DFE & FEGA		78.89%
BU-4DFE & CK+		45.90%	JAFFE & CK+		73.33%
FEGA & BU-4DFE		53.55%	BU-4DFE & JAFFE		55.00%
WSEFEP & BU-4DFE		60.66%	FEGA & JAFFE		73.33%
3 Combined datasets as training set					
WSEFEP, BU-4DFE & CK+	JAFFE	49.18%	FEGA, JAFFE & CK+	WSEFEP	81.67%
FEGA, WSEFEP & CK+		38.25%	CK+, FEGA & BU-4DFE		84.44%
BU-4DFE, FEGA & CK+		49.73%	JAFFE, BU-4DFE & CK+		74.44%
WSEFEP, BU-4DFE & FEGA		59.02%	BU-4DFE, JAFFE & FEGA		81.67%
4 Combined datasets as training set					
WSEFEP, BU-4DFE, FEGA & CK+	JAFFE	46.45%	JAFFE, CK+, BU-4DFE & FEGA	WSEFEP	84.44%

cannot be compared with other works because it is a novel dataset that is presented in this research work.

5.2 Experiment 2. Cross-datasets evaluation

The aim of this experiment is to show how merging information captured with different conditions significantly helps in the network's training. We evaluate the dataset FEGA and our model using a cross-datasets approach. Additionally, we compare the results between the first and the second experiments.

5.2.1 Procedure

Unlike experiment 1 (see Table 9, experiment (a)), where good classification results are obtained using the same dataset to train and test the system, this no longer holds when the training and test datasets are different (cross-datasets approach). In this experiment, our CNN model is evaluated on each dataset (BU-4DFE, CK+, JAFFE, FEGA and WSEFEP) in a cross-datasets approach using 60 epochs each time. The pre-processing that we use (see subsection 3.3) is the same for all datasets. We divide this experiment in three steps: (1) compare the results between the first and the second experiment to classify 6 expressions (see Table 9), (2) explore whether the novel dataset FEGA can be used in a FER system, (3) study whether each dataset adds relevant information in the training in order to ensure diversity in images under different illuminations and backgrounds (see Table 10).

5.2.2 Results and discussion

The results are shown in Tables 9 and 10. Table 9 shows a comparison between both experiments, using the CNN and the image pre-processing. As we can see, when we apply the cross-datasets approach, in general, the results are worse than with subject-independent approach. The reason is that each dataset contains images of people with different ethnicities and ages and different illuminations and backgrounds. For example, the BU-4DFE dataset contains Asian, Black, Latin and White people. The CK+ dataset contains mostly Euro-American people and in a minority, it contains Afro-American and other groups of ethnicities. Both FEGA and WSEFEP datasets contain mostly white people. Finally, the JAFFE dataset is only of Japanese females. Therefore, it is expected that when the training sets are FEGA and WSEFEP datasets, we obtain better results when testing against the CK+ dataset, because it mostly contains Euro-American people. However, CK+ and BU-4DFE obtain best results when tested against the WSEFEP dataset. This suggests that CK+ and BU-4DFE contain an adequate number of white people to be tested with other datasets with white people, as for example, the WSEFEP dataset. On the other hand, the JAFFE dataset is very small for training and the accuracy results when testing against any other dataset are very low.

We also analyze if our new dataset FEGA is a good dataset to train a face expression recognition system. In Table 9, in the cross-datasets experiments (b), the CNN trained with FEGA achieves results of 73.19% and 78.19% against the WSEFEP and CK datasets, because this dataset contains Caucasian people in its majority. In the case of JAFFE and BU-4DFE, we obtain worse results (37.70% and 44.67%) because both datasets contain Asian people and, in the case of BU-4DFE, it also contains Afro-American people. Therefore, FEGA can be considered as a good dataset to train the facial expressions, since it produces good results in

Table 13 Results with different combinations of datasets for the testing set of CK+

Training Set	Test Set	Accuracy
Without combining datasets		
FEGA	CK+	78.19%
BU-4DFE		47.9%
WSEFEP		59.22%
JAFFE		42.97%
2 Combined datasets as training set		
WSEFEP & JAFFE	CK+	55.29%
FEGA & JAFFE		74.54%
FEGA & WSEFEP		81.11%
BU-4DFE & JAFFE		52.65%
FEGA & BU-4DFE		79.84%
WSEFEP & BU-4DFE		73.18%
3 Combined datasets as training set		
WSEFEP, BU-4DFE & JAFFE	CK+	69.71%
FEGA, WSEFEP & JAFFE		77.01%
BU-4DFE, FEGA & JAFFE		81.30%
FEGA, WSEFEP & BU-4DFE		82.66%
4 Combined datasets as training set		
FEGA, JAFFE, BU-4DFE & WSEFEP	CK+	84.76%

Table 14 Comparison of cross-datasets results with related works to classify 6 expressions. Combinations of datasets: (a) FEGA, CK+, BU-4FDE and WSEFEP, (b) FEGA, CK+, BU-4FDE and JAFFE, (c) FEGA, BU-4FDE, JAFFE and WSEFEP, (d) FEGA, CK+, JAFFE and WSEFEP, and (e) JAFFE, CK+, BU-4FDE and WSEFEP, (f) MultiPIE, MMI, DISFA, FERA, SFEW, and FER2013, (g) JAFFE, MMI, RaFD, KDEF, BU3DFE and ARFace, and (h) CK+, MMI, RaFD, KDEF, BU3DFE and ARFace

Training Set	Test Set	Model	Accuracy
Our 4 combined DBs (a) Only BU-4FDE	JAFFE	Our model	46.45%
		Our model	43.17%
6 combined DBs (h) Only CK+		CNN [54]	44.32%
		CNN [28]	38.80%
Our 4 combined DBs (b) Only FEGA	WSEFEP	Our model	84.44%
		Our model	73.89%
Our 4 combined DBs (c) Only FEGA	CK+	Our model	84.76%
		Our model	78.19%
6 averaged DBs (f)		CNN [33]	64.20%
6 combined DBs (g)		CNN [54]	88.58%
Our 4 combined DBs (d) Only CK+ Only CK+	BU-4DFE	Our model	57.68%
		Our model	48.6%
		CNN [28]	45.91%
Our 4 combined DBs (e) Only CK+	FEGA	Our model	55.02%
		Our model	48.6%

a cross-datasets approach, when it is tested with white people, as in the WSEFEP and CK+ datasets.

This suggests that a solution to get successful results is a good combination of different datasets to train the system, which contain all type of ethnicity, age and gender with different illuminations and backgrounds. With the last hypothesis in mind, in Table 10 we show the combinations of the five datasets as training set in order to verify whether each dataset adds important information in the training and ensure a diversity in images regarding ethnicity, age and gender under different illuminations and backgrounds. We test it with each dataset applying k-fold cross-validation with $k = 5$. The high accuracy results in Table 10 show that the CNN distinguishes well between different datasets when it is trained with a dataset containing data with sufficient diversity. Hence, we can claim that each dataset adds an important value in the training. This may not only be due to the diversity in the population, but also due to the different capture conditions of each dataset. Therefore, in the experiment 3, we will evaluate different combinations of datasets for the training set and use the other datasets as testing sets.

Table 15 Comparison between results with four (4 dB) and five (5 dB) combined datasets to classify 6 and 7 expressions

Training Set	BU-4DFE (Test)	FEGA (Test)	JAFFE (Test)	WSEFEP (Test)	CK+ (Test)	Mean
4 dB (6 expressions)	71.67	74.42	60.22	88.89	–	73.80
5 dB (6 expressions)	74.56	76.32	68.45	92.22	94.07	81,12
4 dB (7 expressions)	71.79	72.31	65.36	90.00	–	74.86
5 dB (7 expressions)	71.14	74.10	70.09	91.91	93.23	80.10

5.3 Experiment 3. Different combinations of datasets

In this experiment, we evaluate exhaustively the contributions to learning of different combinations of datasets.

5.3.1 Procedure

Once selected the test set, we need to identify the best datasets for training. One may expect that when using more datasets, we would achieve better results in learning. However, we need new information to learn, therefore, the inclusion of a new dataset will be beneficial when new cases with new information is added. Consequently, we need to determine which datasets contribute more to the test set selected.

To analyze the accuracy when combining different datasets, we create all subset combinations of 5 datasets to train the CNN. Then, we compare the results to study the importance of including different datasets.

5.3.2 Results

The results for different combinations of datasets are presented in Tables 11, 12 and 13. Table 11 shows the accuracy when testing with BU-4DFE and FEGA. Table 12 presents the results testing with JAFFE and WSEFEP and finally, Table 13 shows the results for the CK+ dataset. We show the best results in bold.

In general, four combined datasets achieve better results in most cases. In Table 14, we show results with the following combinations of datasets for training: (1) FEGA, CK+, BU-4FDE and WSEFEP, (2) FEGA, CK+, BU-4FDE and JAFFE, (3) FEGA, BU-4FDE, JAFFE and WSEFEP, (4) FEGA, CK+, JAFFE and WSEFEP, and (5) JAFFE, CK+, BU-4FDE and WSEFEP. These results have been compared with other related works [28, 33, 54] that use CNNs of six classes (one for each facial expression). It can be seen that a good combination of training datasets improves the results. Our results are better in most of cases, only the results presented in [54] are better when testing with CK+. Up to our knowledge, only work [54] combine several datasets. In [33], they train with MultiPIE, MMI, DISFA, FERA, SFEW, and FER2013 datasets separately, use CK+ as testing set and averaged the accuracy results with this testing set (Table 14 (f)). And in [54], they combine JAFFE, MMI, RaFD, KDEF, BU3DFE and ARFace datasets to test with CK+ (Table 14 (g)) and combine CK+, MMI, RaFD, KDEF, BU3DFE and ARFace datasets to test with JAFFE (Table 14 (h)). Unfortunately, we have not found works which have been tested with the WSEFEP dataset in a cross-datasets evaluation scenario. And obviously the FEGA dataset has been tested for the first time.

Table 16 Comparison between results with five combined datasets (5 dB) and the results of the subject-independent experiment (1 dB) for the classification of 6 expressions

Training Set	BU-4DFE (Test)	FEGA (Test)	JAFFE (Test)	WSEFEP (Test)	CK+ (Test)	Mean (Test)
1 dB (Subject-indep. experiment) (6 expressions)	73.58	72.61	56.22	87.22	93.15	76.48
5 dB (combined datasets) (6 expressions)	74.56	76.32	68.45	92.22	94.07	81.12

Table 17 Results of the five (5 dB) combined datasets to classify 7 expressions using different architectures and the same image pre-processing

Model	BU-4DFE	FEGA	JAFFE	WSEFEP	CK+	Mean
AlexNet [22]	71.21	70.32	67.56	90.95	91.76	78.36
Burkert et al. [6]	57.09	55.61	45.66	70.48	83.44	62.46
Khorrami et al. [20]	72.31	73.62	68.43	91.43	90.82	79.32
Song et al. [45]	72.67	69.57	64.33	88.09	88.36	76.60
Our Model	71.14	74.1	70.09	91.91	93.23	80.10

Therefore, in summary, the combination of several datasets to train the system improves the results according to Tables 11 and 14. Based on these findings, we detail the results obtained when we train with the four datasets of the case (c) of Table 14, and when we train with the five datasets (see Table 15). We have performed k-fold cross-validation using $k = 5$ to classify both six and seven expressions using our CNN. That is, we separate these datasets (4 combined DBs and 5 combined DBs) in 5 blocks both in the training set and in the test set. For example, we train with blocks 1, 2, 3 and 4 (with data augmentation), and we test with block 1 (without data augmentation). Each block consists of a fifth part of all the combined datasets (BU-4DFE, CK+, JAFFE, FEGA and WSEFEP).

5.3.3 Discussion

To maximize the success of a neural network model R using N datasets, we define the set of datasets used for learning as $D = \{D_1, D_2, \dots, D_N\}$, where N corresponds to the number of available datasets. To select the best combination of datasets, we need to test all possible combinations for each subset of D (see Tables 11, 12 and 13), except for the empty set \emptyset . Each table is divided into four groups of combinations (without combining datasets, 2 combined datasets, 3 combined datasets and 4 combined datasets). The number of combinations to test is $card(D) - 1 = 2^N - 1$. For example, with four datasets, we need to train and evaluate the network $2^4 - 1 = 15$ times. If possible, we would like to reduce the number of combinations to use the ones that contribute more, as e.g., it would be unpracticable to use 30 datasets.

We define the function $f(R, T)$, where T is the Test dataset, and R is a neural network model. The function f returns the subset of D that achieves the best accuracy. We define the function $f_k(R, T)$ for k in $[1, N]$. This function returns the subset of D using k elements that achieve the best accuracy. For example, $f_3(\text{Our Model}, \text{JAFFE})$ would return $\{\text{WSEFEP}, \text{BU-4DFE}, \text{FEGA}\}$. Note that $f(R, T)$ can be computed from all $f_k(R, T)$ by comparison. As we can see in Tables 11, 12 and 13, for each test set, we train 15 different combinations. This means a high computational cost when the number of datasets augments.

In each group of each table (Tables 11, 12 and 13), the best result is highlighted in bold. Note that the best result of each group contains the dataset of the previous combination. This gives a hint on how to reduce the number of combinations of datasets that need to be checked in order to obtain the optimum result.

We propose an iterative procedure in which k datasets are used at each step. If we denote B_k as the optimum set of datasets used at step k ($B_k = f_k(R, T)$), B_k is defined as $B_{k-1} \cup D_{k^*}$, where D_{k^*} is the dataset in $D - B_{k-1}$ which maximizes the accuracy of the combined dataset $B_{k-1} \cup D_j$, for $D_j \in D - B_{k-1}$. By definition $B_0 = \emptyset$. Under this premise, we can obtain the best training with a computational cost N^2 . An example for $N = 4$ is shown in (2),

Table 18 Results of the test FE-Test (6 expressions) using different datasets as training set

Training set	Test set	Accuracy
BU-4FDE	FE-Test	51.72%
FEGA		60.67%
JAFFE		31.56%
WSEFEP		61.78%
CK+		55.72%
5 dB		73.05%

$$\begin{aligned}
 B_0 &= \emptyset \\
 B_1 &= \{D_3\} \\
 B_2 &= \{D_3, D_1\} \\
 B_3 &= \{D_3, D_1, D_4\} \\
 B_4 &= \{D_3, D_1, D_4, D_2\} = D
 \end{aligned} \tag{2}$$

where D_j is the added dataset that maximizes the function $f(R, T)$ until $k = N$, $B_N = D$. With this procedure, we reduce the number of trainings needed to obtain B_N to:

$$\sum_{i=1}^N i = \frac{N*(N+1)}{2} \tag{3}$$

For example, if $N = 100$, instead of training the neural network R with $2^{100}-1$ combinations of datasets, we would reduce the number of trainings to 5050 combinations (4).

$$\sum_{i=1}^{100} i = \frac{100*(100+1)}{2} = 5050 \tag{4}$$

One aspect to keep in mind is that adding a new dataset does not always improve the results. So, we must determine the value of k for which the function, $f(R_{B_k}, T)$, is maximum.

Therefore, in order to know the test accuracy in each dataset, we separate the images of this test block in each dataset. Results improve in all cases when training with five datasets, except when testing with BU-4FDE, which obtains similar results in the case of seven expressions. On the whole, we improve the results to 80.1% in the test accuracy to classify seven expressions and 81.12% to classify six expressions. In addition, we also improve the accuracy achieved when only one dataset is used for training and testing (see Table 16).

5.4 Experiment 4. Comparison of our system with other architectures

The aim of this experiment is to compare the performance of our system regarding other existing architectures.

Table 19 Results of the test FE-Test (6 expressions) using different CNN

Model	Training set	Test set	Accuracy	σ
Burkert et al. [6]	5 dB	FE-Test	56.83%	4.20
Khorrami et al. [20]			72.99%	1.17
Song et al. [45]			71.44%	1.54
AlexNet [22]			72.77%	1.81
Our Model			73.05%	0.94

5.4.1 Procedure

To verify the proper functionality of the CNN of our system, we have compared it with several CNNs [6, 20, 22, 45], using the five combined datasets for training and applying our image pre-processing. We implemented the architectures [6, 20, 45] following the descriptions in the corresponding papers. These models were specifically created for the FER task. Moreover, we also tested the performance of the well-known AlexNet [22] network architecture, which is available in the Caffe framework.

5.4.2 Results and discussion

The results from each CNN are shown in Table 17. Our CNN shows the best results in most cases, except for the BU-4DFE test dataset, where the Song et al. network [45] achieves better results. Therefore, we can claim that our CNN is competitive with respect to other existing CNNs and that works well for FER.

5.5 Experiment 5. Evaluation of an unknown test dataset

The aim of this experiment is to analyze the accuracy of our system recognizing 6 facial expressions using an unknown dataset (FE-test).

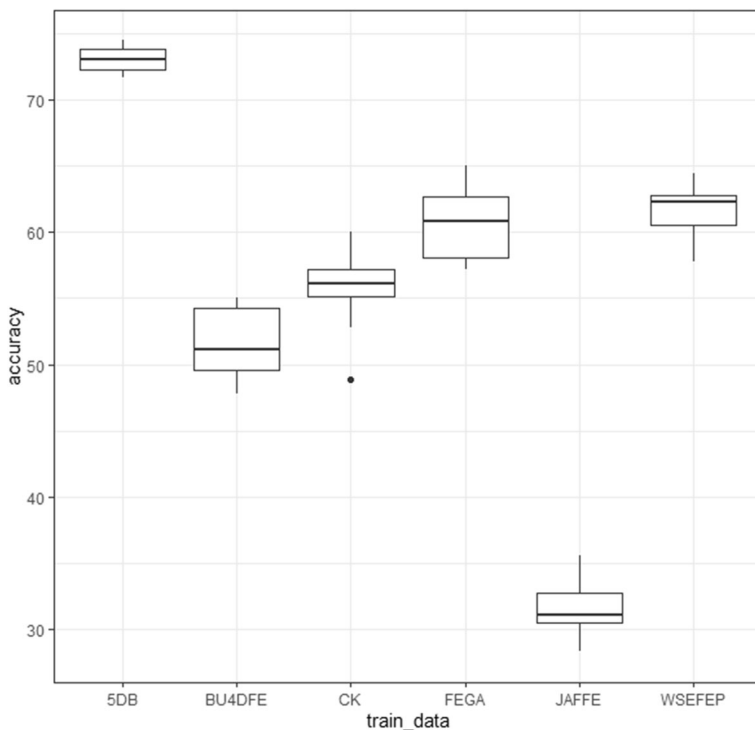


Fig. 8 Box plot for accuracy percentage results training neural networks using 5 databases, and each one individually. Accuracy has been computed using FE-Test. The neural network is trained 10 times for each population (training dataset), so we have a total of 60 samples

5.5.1 Procedure

We employ our CNN and image pre-processing steps to classify expressions of the FE-Test dataset. We first study the performance of the system to discriminate between 6 expressions. The system is trained in two ways: with each of the five datasets separately available from previous Sections (WSEFEP, BU-4DFE, CK+, JAFFE, FEGA), and with the five combined datasets together (5 dB). Each training set is run for 10 times and the results show the mean of these 10 runs. Second, we select the best training set. We train it with different CNNs [6, 20, 22, 45] and compare them with the result of our CNN.

5.5.2 Results and discussion

Table 18 shows that the best results are obtained with the combination of five datasets. As we see, we improve the results up to a 41.49% (the worst result is achieved with the JAFFE as training dataset, which gets 31.56%, while using the combination of DBs for training results in an accuracy of 73.05%). And we improve a 11.27% the best result obtained with only one dataset.

Besides our model, we use different models to perform the experiment with the training set that has obtained the best result in Table 18 (5 dB). Since the combination of the five datasets obtains a considerable improvement in the results, we compare this result (73.05%) with the other models described in the Experiment 4. This comparison is shown in Table 19. For each model, we show the mean in the accuracy and the standard deviation. For this set of architectures, the best result is obtained with our model, with which we achieve the highest accuracy and the lowest standard deviation.

To confirm the improvement when using the cross-dataset approach, we carry out a statistical analysis [55] to assess if the use of five datasets improves the results compared with a single dataset for training.

We check statistically if the use of 5 databases improves the results compared to training with a single database. The experiment carried out trains the proposed CNN using the 5 databases, CK +, BU4DFE, FEGA, JAFFE, WSEFEP (6 populations), and measures the accuracy obtained in an independent database: FE-Test. For each case, we train 10 times, having a total of 60 samples (10 per training dataset).

To confirm the hypothesis, we perform an ANOVA test, where the null hypothesis is: the equality of the means, and the alternative hypothesis is: there are at least two populations with different mean. To be able to apply ANOVA, it must be fulfilled that populations come from the same distribution, in our case that they follow a normal distribution, and they fulfil homoscedasticity.

We perform the Shapiro-Wilk test to verify normality, and the Barlett test to verify homoscedasticity. The Shapiro-Wilk test yields the following values of the statistic W: 0.91962, 0.92177, 0.89428, 0.93899, 0.96621 and 0.90559 (with p values 0.332, 0.372, 0.1894, 0.5418, 0.8537, 0.252) so we assume, without any doubt, that the six populations come from normal distributions. In Bartlett's test, the statistic obtained follows a chi-square distribution of 5 degrees of freedom (having 6 populations), resulting in $X^2 = 10.764$, and a p value = 0.05626. Being able to assume homoscedasticity.

The ANOVA test clearly rejects the null hypothesis, accepting that the means are not equal (at least two means are different), with a p value $< 2 \cdot 10^{-16}$. Performing the pairwise comparison

Table 20 Experts agreement level with a sample of 1000 images from AffectNet

Agreement	Number images
4	362
3–1	270
2–2	58
2–1–1	243
1–1–1–1	67

using Tukey HSD, it is obtained that the only equality occurs between WSEFEP-FEGA with p value 0.8958134. For the rest of pairs we obtained p-values lower than 0.005.

Statistical analysis confirms that using all five datasets improves performance over using only one dataset. Figure 8 shows the box plot for the 6 populations.

6 Human performance evaluation

In this Section we first analyze and reflect on the subjectivity inherent to human nature when labelling facial expression images. Then, we present an experiment to study how close the results obtained by our system are to those obtained by human perception. Figure 7(c) summarizes the experiments performed in this step. We carry out two evaluations with the FE-Test dataset: in the first test, humans classify the images, while in the second test, we use the CNN. The task is to classify the images in the six basic emotions (anger, disgust, fear, happiness, sadness and surprise) plus the neutral one. As already commented, the FE-test dataset contains facial expression images “in the wild” from the Internet validated by the research team with a wide range of backgrounds and lightings. Therefore, recognition by the CNN is expected to be more difficult due to these conditions. Further, the FE-test has not been previously used in any other experiment; therefore, we will also observe how the system performs to new data.

6.1 How subjective is facial expression recognition by human experts?

To analyze the agreement level in image labelling, we evaluate four experts labelling images obtained from web scraping, specifically, 1000 randomly selected images with no duplicates from the AffectNet dataset. To build the final sample, we generate 100 samples to select the one with the highest similarity regarding proportions of classes in the dataset using the chi-square similarity measure. The number of samples used for each class and the number of samples provided by AffectNet were: Neutral 182/75374, Happy 320/134915, Sad 57/25959,

Table 21 Kappa-Cohen index for the experts. The highest agreement is between experts E2 and E3. The lowest agreement is between experts EA (AffectNet) and E1

	E1	E2	E3
EA	0.4260	0.4315	0.4620
E1		0.4374	0.4736
E2			0.5299

Table 22 Number of images for each class labelled by each expert (EA, E1, E2 y E3). Column C is the number of images in which the four experts agree. And the last column is the percentage of agreement regarding the minimum between the four evaluators

	EA	E1	E2	E3	C	%
Neutral	182	127	170	337	49	24.02
Happy	320	269	412	315	221	67.17
Sad	57	54	78	61	17	27.20
Surprise	33	28	26	24	5	18.02
Fear	18	23	17	18	2	10.53
Disgust	11	12	22	17	1	6.45
Anger	59	32	43	32	7	16.87
Contempt	10	11	18	10	0	0.00
None	78	159	118	54	4	3.91
Uncertain	23	185	24	40	1	1.47
Non-face	209	100	72	92	55	46.51

Surprise 33/14590, Fear 18/6878, Disgust 11/4303, Anger 59/25382, Contempt 10/4250, None 78/33588, Uncertain 23/12145 and Non-Face 209/82915. For each class, the number of images is over five and we can use a chi-square contrast without correction.

Four experts participated: the expert or experts that labelled the AffectNet dataset (EA) and the three experts that achieved more accurate results in the experiment presented in [39], which were: 76.7% (E10), 75.1% (E4) y 73.2% (E6) correspondently. In this work these experts are labeled as E1, E2 and E3. The aim of the study is to determine the agreement level among experts, and therefore, we do not consider having a ground truth. The level of agreement is summarized in Table 20.

The agreement level is measured with the Kappa-Cohen index (see Table 21). To consider a reliable agreement between two experts, the index should be over 0.60 in the case of two classes, although this can be decreased when there is a high number of classes (11 expressions) [24]. In all cases, we observe a moderate agreement ($0.41 \leq k \leq 0.60$).

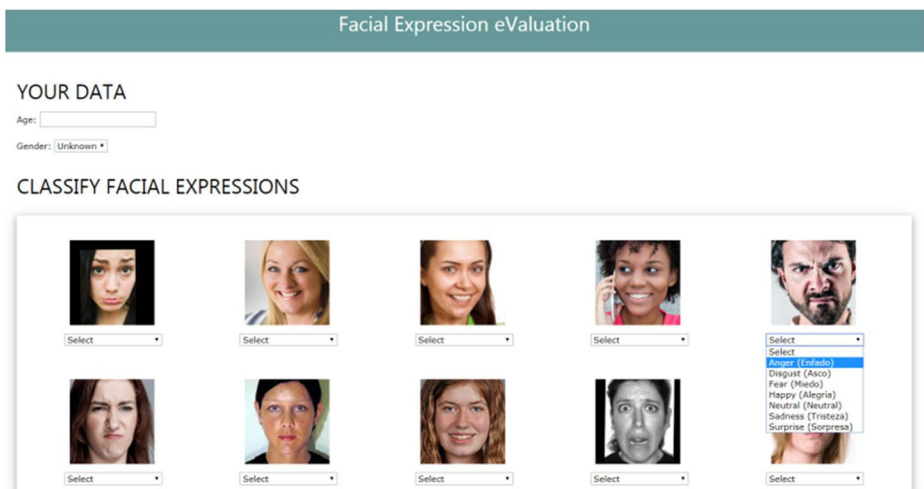


Fig. 9 The created web page for the experiment of FER by humans

Table 23 Confusion Matrix from human assessment (7 expressions). Results of the FE-Test dataset using a cross-datasets approach

	AN	DI	FE	HA	NE	SA	SU	Total
AN	329	21	5	2	3	2	3	90.14%
DI	23	303	14	3	1	14	7	83.01%
FE	7	22	243	0	1	5	88	66.39%
HA	1	2	2	331	12	2	2	94.03%
NE	6	4	5	13	331	5	0	90.93%
SA	7	12	11	2	45	276	7	76.67%
SU	5	4	13	29	7	1	299	83.52%

When humans label expressions, there is a level of subjectivity due to the human nature. In Table 22, we show that Happiness is one of the easiest expressions to classify (observing the agreement level). We highlight that all experts agree only in 36.2% of the images, which can be due to the subtlety of some expressions, being many of them ambiguous. Regarding the classes, in descending order of level of agreement, we have: Happiness, Non-face, Sadness, Neutral, Surprise, Anger, Fear, Disgust, None, Uncertain and Contempt. Table 22 summarizes the results, the first four columns (EA, E1, E2 and E3) count the number of images labelled in a specific class. Column C shows the number of cases in which the four experts agree on the label. Finally, the agreement level is calculated in relation to the average number of labels of an expression. The classification of expressions for a human is very subjective and using the labelling of such expressions as ground truth leads to reproducing the subjectivity of the human expert. However, the use of datasets where professional actors pose a facial expression (even if the person is not feeling that real emotion) is a more reliable data source.

6.2 Facial expression recognition by humans

The aim of this experiment is to study the human perception to recognize facial expressions using the FE-test dataset, taking into account the subjectivity of humans studied in the previous subsection.

6.2.1 Participants

253 unpaid, voluntary participants (27.27% women) were involved in the study with ages ranging from 18 to 66 years. Participants were recruited via social media.

Table 24 Confusion Matrix from our system (7 expressions). Results of the FE-Test dataset using the cross-datasets approach

	AN	DI	FE	HA	NE	SA	SU	Total
AN	73	55	2	4	3	2	11	48.67%
DI	22	117	7	0	0	4	0	78.00%
FE	6	5	82	0	3	5	49	54.67%
HA	0	1	2	147	0	0	0	98.00%
NE	31	1	5	5	101	6	1	67.33%
SA	29	24	6	2	9	74	6	49.33%
SU	1	2	10	4	0	4	129	86.00%

6.2.2 Task

The task to carry out was the classification of 10 random images into one of the seven emotions: AN (angry), DI (Disgust), FE (Fear), HA (Happiness), SA (Sadness), SU (Surprise) and NE (Neutral)).

6.2.3 Procedure

Participants were informed on the goal of the research and the task to carry out. They received a link to a webpage where 10 random images from the FE-test dataset were shown with a drop-down list including the seven emotions (see Fig. 9). Further, we gathered demographical data regarding age and gender.

6.2.4 Results and discussion

Table 23 present the results obtained by the 253 participants in form of the confusion matrix with an average value of 83.53% for all classes. On the one hand, we can observe that there are some expressions that are confused with others such as Sadness and Fear, which are often confused with Neutral and Surprise, respectively. On the other hand, Happiness is the clearest to distinguish, and most of the participants recognize it easily.

6.3 Facial expression recognition by our system

The aim of this experiment is to analyze the accuracy of our system recognizing 7 facial expressions in FE-test, an unknown dataset, and compare these results with the ones obtained by humans.

6.3.1 Procedure

We use our CNN and image pre-processing steps to classify the 7 expressions of the FE-Test dataset, in order to compare it with the results obtained by humans. Each training set is run for 10 times and the results show the mean of these 10 runs.

6.3.2 Results and discussion

Based on the results of the Tables 22 and 23, we use the combination of five datasets as training set using our CNN and pre-processing step to evaluate the recognition of 7 expressions.

We use the FE-Test as testing set; thus, this is also a cross-datasets evaluation. The confusion matrix is shown in Table 24, where we have obtained a total average of 68.86% in the accuracy. The higher accuracy is achieved with the Happiness and Surprise expressions, with the CNN performing better than humans. Further, both experiments obtained the worst results with Sadness and Fear, although humans perform better than the CNN. Results are also higher for humans in recognizing the Neutral, Angry and Disgust emotions. Finally, we can also see a correlation between the experiments, especially in the recognition of Angry, Disgust and Fear, which are usually confused with Disgust, Angry and Surprise, respectively.

Interestingly, these mistakes are done both by humans and machines, that is, both perform similar misclassifications.

7 Conclusions

FER is a difficult task both for human and technology. In the case of automatic FER using CNNs, issues such as the model or the datasets used for training and testing are fundamental to achieve accurate results. This work presented an extensive evaluation of a CNN using both single and cross-datasets approaches, considering initial steps such as the selection of datasets to include, image pre-processing or even a comparison with human perception. Up to our knowledge, this is the most extensive experimental study to date in cross-dataset FER using CNNs, since most previous studies in the literature only employ one dataset for testing.

We first defined a protocol to select and work with different datasets and create a homogenized dataset with data augmentation. As far as we know, there is no systematic protocol to carry out this task. We listed a set of steps to follow highlighting restrictions and considerations. Further, we mentioned some methods to use in the different steps of the protocol in subsection 3.3: Image Pre-Processing and Data Augmentation. When selecting the dataset to use, we highlighted the problems when using facial expressions datasets built via web scraping regarding duplicated images or even with non-face images.

Then, we presented empirical results of an exhaustive evaluation (using single and cross-datasets approaches) analyzing the relevance of a pre-processing step to improve the performance despite the intrinsic complexity of a CNN and how merging information from diverse datasets significantly helps in the network's training task. Our study shows that each dataset adds an important value in the training, because each of them has been captured in different conditions and contains people from different ethnicities and ages. Therefore, not only the quantity is important to train the data with CNN, but also the variety of information. Thus, the combination of these datasets into one single training dataset, using our image pre-processing steps to unify them significantly improves the results with respect to using only one dataset for training. Further, we move forward to apply automatic FER in the wild. Our experiments also show that our system outperforms other proposed solutions in the literature (see Table 17) and presents good accuracy results in real world situations. We highlight the result of around 70% in accuracy using the cross-datasets approach when the test set comes from a never-seen-before dataset (evaluations carried out with FE-test). Finally, we compared human versus our CNN classification, observing similar results and similar misclassifications errors.

As future work, we intend to refine our system with more datasets, in addition to studying the pre-processing step for color images. We also plan to extend this study to build a metric to score learning influence of the datasets and its bias.

Acknowledgements We want to thank to the University of Cagliari for the collaboration in the experiments of this work.

Funding Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature. This work has been supported by the project PID2019-104829RA-I00/MCIN/AEI/<https://doi.org/10.13039/501100011033>, EXPLainable Artificial INtelligence systems for health and well-beING (EXPLAINING), and the project RTI2018-096986-B-C31 (MINECO/AEI/ERDF, EU), Design of pervasive gaming experiences for intergenerational social and emotional well-being (PERGAMEX).

Declarations

Conflict of interest The authors have no conflicts of interest to declare that are relevant to the content of this article.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

1. Abdullah SMSA, Ameen SYA, Sadeeq MAM, Zeebaree S (2021) Multimodal emotion recognition using deep learning. *J Appl Sci Technol Trends* 2:52–58. <https://doi.org/10.38094/jastt20291>
2. Ali A, Zhu Y, Zakarya M (2021) A data aggregation based approach to exploit dynamic spatio-temporal correlations for citywide crowd flows prediction in fog computing. *Multimed Tools Appl* 80:31401–31433. <https://doi.org/10.1007/s11042-020-10486-4>
3. Barreto AM (2017) Application of facial expression studies on the field of marketing. *Emotional Expression: The Brain and the Face Porto: FEELab Science Books* 163–189
4. Barsoum E, Zhang C, Ferrer CC, Zhang Z (2016) Training deep networks for facial expression recognition with crowd-sourced label distribution. In: *ICMI 2016 - proceedings of the 18th ACM international conference on multimodal interaction*
5. Berrar D (2018) Cross-validation. In: *encyclopedia of bioinformatics and computational biology: ABC of bioinformatics*
6. Burkert P, Trier F, Afzal MZ et al (2015) DeXpression: Deep Convolutional Neural Network for Expression Recognition
7. Chen Z, Zhang Y, Wu C, Ran B (2019) Understanding individualization driving states via latent Dirichlet allocation model. *IEEE Intell Transp Syst Mag* 11:41–53. <https://doi.org/10.1109/ITS.2019.2903525>
8. Ekman P (1977) Facial-expression. In: Siegman A, Feldstein S (eds) *Nonverbal communication and behavior*. Lawrence Erlbaum association, New Jersey, pp 97–126
9. Fischer A, Lafrance M (2015) What drives the smile and the tear: why women are more emotionally expressive than men. *Emot Rev* 7:22–29. <https://doi.org/10.1177/1754073914544406>
10. Gilyazev RA, Turdakov DY (2018) Active learning and crowdsourcing: a survey of optimization methods for data labeling. *Program Comput Softw* 44:476–491. <https://doi.org/10.1134/S0361768818060142>
11. Giryes R, Sapiro G, Bronstein AM (2016) Deep neural networks with random Gaussian weights: a universal classification strategy? *IEEE Trans Signal Process* 64:3444–3457. <https://doi.org/10.1109/TSP.2016.2546221>
12. Glorot X, Bengio Y (2010) Understanding the difficulty of training deep feedforward neural networks. *J Mach Learn Res*
13. Grabowski K, Rynkiewicz A, Lassalle A et al (2019) Emotional expression in psychiatric conditions: new technology for clinicians. *Psychiatry Clin Neurosci* 73:50–62. <https://doi.org/10.1111/pcn.12799>
14. Han B, Yun W-H, Yoo J-H, Kim WH (2020) Toward unbiased facial expression recognition in the wild via cross-dataset adaptation. *IEEE Access* 8:159172–159181. <https://doi.org/10.1109/ACCESS.2020.3018738>
15. Hinton GE, Osindero S, Teh YW (2006) A fast learning algorithm for deep belief nets. *Neural Comput* 18. <https://doi.org/10.1162/neco.2006.18.7.1527>
16. Jain N, Kumar S, Kumar A et al (2018) Hybrid deep neural networks for face emotion recognition. *Pattern Recogn Lett* 115:101–106. <https://doi.org/10.1016/j.patrec.2018.04.010>
17. Jain DK, Shamsolmoali P, Sehdev P (2019) Extended deep neural network for facial emotion recognition. *Pattern Recogn Lett* 120:69–74. <https://doi.org/10.1016/j.patrec.2019.01.008>
18. James G, Witten D, Hastie T, Tibshirani R (2013) *An introduction to statistical learning with applications in R*
19. Kazemi V, Sullivan J (2014) One millisecond face alignment with an ensemble of regression trees. In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*
20. Khorrami P, Le Paine T, Huang TS (2015) Do deep neural networks learn facial action units when doing expression recognition? In: *Proceedings of the IEEE International Conference on Computer Vision*

21. Kingma DP, Ba JL (2015) Adam: a method for stochastic optimization. In: 3rd international conference on learning representations, ICLR 2015 - conference track proceedings
22. Krizhevsky A, Sutskever I, Hinton GE (2012) ImageNet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems
23. Küntzler T, Höfling TTA, Alpers GW (2021) Automatic facial expression recognition in standardized and non-standardized emotional expressions. *Front Psychol* 12:1086. <https://doi.org/10.3389/fpsyg.2021.627561>
24. Landis JR, Koch GG (1977) The measurement of observer agreement for categorical data. *Biometrics* 33: 159. <https://doi.org/10.2307/2529310>
25. Li Y, Fan C, Li Y et al (2018) Improving deep neural network with multiple parametric exponential linear units. *Neurocomputing* 301:11–24. <https://doi.org/10.1016/j.neucom.2018.01.084>
26. Lisani JL, Ramis S, Perales FJ (2017) A contrario detection of faces: a case example. *SIAM J Imaging Sci* 10:2091–2118. <https://doi.org/10.1137/17M1118774>
27. Liu S, Tang X, Wang D (2020) Facial expression recognition based on sobel operator and improved CNN-SVM. In: 2020 3rd IEEE international conference on information communication and signal processing, ICICSP 2020
28. Lopes AT, de Aguiar E, De Souza AF, Oliveira-Santos T (2017) Facial expression recognition with convolutional neural networks: coping with few data and the training sample order. *Pattern Recogn* 61: 610–628. <https://doi.org/10.1016/j.patcog.2016.07.026>
29. Lucey P, Cohn JF, Kanade T et al (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, CVPRW 2010
30. Lyons M, Kamachi M, Gyoba J (2017) Japanese female facial expression (JAFFE) database. Available: <http://www.kasrl.org/jaffe.html>
31. Malatesta CZ, Izard CE, Culver C, Nicolich M (1987) Emotion communication skills in young, middle-aged, and older women. *Psychol Aging* 2:193–203. <https://doi.org/10.1037/0882-7974.2.2.193>
32. Medjdjen S, Ahmed N, Lataifeh M (2020) Adaptive user interface design and analysis using emotion recognition through facial expressions and body posture from an RGB-D sensor. *PLoS One* 15:e0235908. <https://doi.org/10.1371/journal.pone.0235908>
33. 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 2016
34. Mollahosseini A, Hasani B, Mahoor MH (2017) AffectNet: a database for facial expression, valence, and arousal computing in the wild. *IEEE Trans Affect Comput* 10. <https://doi.org/10.1109/TAFFC.2017.2740923>
35. Najah G, Şengül G (2017) Emotion estimation from facial images. Turkey
36. Olszanowski M, Pochwatko G, Kuklinski K et al (2014) Warsaw set of emotional facial expression pictures: a validation study of facial display photographs. *Front Psychol* 5. <https://doi.org/10.3389/fpsyg.2014.01516>
37. Owusu E, Zhan Y, Mao QR (2014) A neural-AdaBoost based facial expression recognition system. *Expert Syst Appl* 41:3383–3390. <https://doi.org/10.1016/j.eswa.2013.11.041>
38. Poursaberi A, Noubari HA, Gavrilova M, Yanushkevich SN (2012) Gauss-Laguerre wavelet textural feature fusion with geometrical information for facial expression identification. *Eurasip J Image Video Process* 17(2012). <https://doi.org/10.1186/1687-5281-2012-17>
39. Ramis S, Buades JM, Perales FJ (2020) Using a social robot to evaluate facial expressions in the wild. *Sensors (Switzerland)* 20. <https://doi.org/10.3390/s20236716>
40. Sagonas C, Tzimiropoulos G, Zafeiriou S, Pantic M (2013) 300 faces-in-the-wild challenge: the first facial landmark localization challenge. In: Proceedings of the IEEE International Conference on Computer Vision, pp 397–403
41. Sajjanhar A, Wu Z, Wen Q (2019) Deep learning models for facial expression recognition. In: 2018 international conference on digital image computing: techniques and applications, DICTA 2018
42. Sebe N, Lew MS, Sun Y et al (2007) Authentic facial expression analysis. *Image Vis Comput* 25:1856–1863. <https://doi.org/10.1016/j.imavis.2005.12.021>
43. Shamsheerband S, Fathi M, Dehzangi A et al (2021) A review on deep learning approaches in healthcare systems: taxonomies, challenges, and open issues. *J Biomed Inform* 113:103627. <https://doi.org/10.1016/j.jbi.2020.103627>
44. Siddiqi MH, Ali R, Sattar A et al (2014) Depth camera-based facial expression recognition system using multilayer scheme. *IETE Technical Review (Institution of Electronics and Telecommunication Engineers, India)* 31:277–286. <https://doi.org/10.1080/02564602.2014.944588>
45. Song I, Kim HJ, Jeon PB (2014) Deep learning for real-time robust facial expression recognition on a smartphone. In: Digest of Technical Papers - IEEE International Conference on Consumer Electronics
46. Susskind JM, Littlewort G, Bartlett MS et al (2007) Human and computer recognition of facial expressions of emotion. *Neuropsychologia* 45:152–162. <https://doi.org/10.1016/j.neuropsychologia.2006.05.001>
47. Taner Eskil M, Benli KS (2014) Facial expression recognition based on anatomy. *Comput Vis Image Underst* 119:1–14. <https://doi.org/10.1016/j.cviu.2013.11.002>

48. Trujillo L, Olague G, Hammoud R, Hernandez B (2005) Automatic feature localization in thermal images for facial expression recognition. In: Proceedings - 2005 IEEE computer society conference on computer vision and pattern recognition, CVPR 2005-Workshops. IEEE, pp 14–14
49. Valstar MF, Pantic M (2010) Induced disgust, happiness and surprise: an addition to the MMI facial expression database
50. Valstar MF, Almaev T, Girard JM et al (2015) FERA 2015 - second facial expression recognition and analysis challenge. In: 2015 11th IEEE international conference and workshops on automatic face and gesture recognition, FG 2015
51. Wang Z, Bovik AC, Sheikh HR, Simoncelli EP (2004) Image quality assessment: from error visibility to structural similarity. *IEEE Trans Image Process* 13:600–612. <https://doi.org/10.1109/TIP.2003.819861>
52. Wen G, Hou Z, Li H et al (2017) Ensemble of Deep Neural Networks with Probability-Based Fusion for Facial Expression Recognition. *Cogn Comput* 9:597–610. <https://doi.org/10.1007/s12559-017-9472-6>
53. Yin L, Wei X, Sun Y et al (2006) A 3D facial expression database for facial behavior research. In: FGR 2006: proceedings of the 7th international conference on automatic face and gesture recognition
54. Zavarez MV, Berriel RF, Oliveira-Santos T (2017) Cross-database facial expression recognition based on fine-tuned deep convolutional network. In: Proceedings - 30th conference on graphics, patterns and images, SIBGRAPI 2017
55. Zhang Z, Hong W-C (2021) Application of variational mode decomposition and chaotic grey wolf optimizer with support vector regression for forecasting electric loads. *Knowl-Based Syst* 228:107297. <https://doi.org/10.1016/j.knosys.2021.107297>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Silvia Ramis, Ph. D. in Information and Communications Technologies from the UIB (since 2019). She has participated in several projects in the field of Computer Vision, Artificial Intelligence and Human-Robot Interaction. Her research experience focuses on artificial intelligence applied to human-robot interaction, especially in face detection and facial expression recognition.



Jose M. Buades received his degree in Computer Science and his Ph. D. in Computer Science from the University of Balearic Islands. He is currently an Associate Professor at the University of the Balearic Islands. His research interests include computer graphics, computer vision and artificial intelligence.



Francisco J. Perales is a senior researcher in the area of computer graphics and vision, human computer interaction and neuro-rehabilitation. He obtained his BS and MS in computer science at the UAB. He received a Ph. D. degree in Computer Science at the UIB in 1993. He is a and head of the UGIVIA research group at the UIB.



Cristina Manresa-Yee received her degree in Computer Science and her Ph. D. in Computer Science from the University of Balearic Islands. She is currently an Associate Professor at the University of the Balearic Islands. Her research interests include human-computer interaction, computer vision and explainable AI.