



Multichannel convolutional neural network for human emotion recognition from in-the-wild facial expressions

Hadjer Boughanem¹ · Haythem Ghazouani^{1,2} · Walid Barhoumi^{1,2} 

Accepted: 26 September 2022 / Published online: 21 October 2022

© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2022

Abstract

Facial emotions reflect the person's moods and show the human affective state that is correlative with non-verbal intentions and behaviors. Despite the advances on computer vision techniques, capturing automatically the facial expressions in-the-wild remains a very difficult task. In this context, we propose a multichannel convolutional neural network based on the quality and the strengths of three well-known pre-trained models, namely VGG19, GoogleNet, and ResNet101. Indeed, the complementarity of the features extracted from the three models is exploited in order to form a more robust feature vector. During the training phase, a freezing weight is applied for each architecture. Then, the layers containing the most relevant information are marked, and the final feature descriptor for emotion prediction is thereafter defined by concatenating the obtained vectors. In fact, the three architectures have showed their efficiency severally in term of emotion recognition, and notably they do not err in the same images. The final vector, obtained by concatenating the features extracted from the different models, is fed to a support vector machine classifier in order to predict the final emotions. Extensive experiments have been conducted on four challenging datasets (JAFFE, CK+, FER2013 and SFEW_2.0) covering in-the-wild as well as in-the-laboratory facial expressions. The obtained results show that the suggested method is not only more accurate compared to each pre-trained CNN model but it also outperforms relevant state-of-the-art methods.

Keywords Deep features · Multichannel CNN · In-The-Wild emotion recognition · Human emotion recognition · Feature concatenation

1 Introduction

Most people believe knowing a great deal about their own emotions, nevertheless psychologists face difficulties in having a consensus about the nature and the working mechanisms of emotions [1]. Emotions, which are relatively brief, are fundamental human features playing important roles in social communication and effecting all social phenomena [2]. These emotions allow the observer to infer the emotional

states as well as the intentions of others, which make it possible to anticipate their gestures and regulate his own behaviors accordingly. Emotions are evinced by different reactions such as psychological reactions change in tone voice, palpitations, heat, accelerated pulse gestural expressions and facial expressions. However, defining the human emotion is not simple, and the interest of many of researchers are aroused by the complexity that emotions carry [3]. Darwin has emphasized that emotion is a response to the environment [4], while Dam et al. [5] have defined the emotion as a reaction to an event which appears suddenly, without lasting long. Several existing works have the unanimous goal of classifying the input emotion into one of the seven basic emotion classes (happiness, sadness, neutrality, disgust, fear, surprise, and anger). These works just differ in the modalities used [6] and the supports treated from which the features and the information are extracted in order to be able to predict the emotions [7]. Among the relevant modalities, facial expressions are one of the most popular [8], due to several reasons. They are visible, they contain many useful features for emotion

✉ Walid Barhoumi
walid.barhoumi@enicarthage.rnu.tn

¹ Institut Supérieur d'Informatique d'El Manar, Research Team on Intelligent Systems in Imaging and Artificial Vision (SIIVA), LR16ES06 Laboratoire de recherche en Informatique, Modélisation et Traitement de l'Information et de la Connaissance (LIMTIC), Université de Tunis El Manar, 2 Rue Abou Rayhane Bayrouni, 2080 Ariana, Tunisia

² Ecole Nationale d'Ingénieurs de Carthage, Université de Carthage, 45 Rue des Entrepreneurs, 2035 Tunis-Carthage, Tunisia

recognition, and it is relatively easy to collect a large dataset of face images [9]. It is worth mentioned that image datasets designed under controlled laboratory conditions are more available than those designed under uncontrolled (in-the-wild) conditions. Among them, we point out the most widely used ones, such as the Japanese Female Facial Expression (JAFFE) dataset [10], the Cohn-Kanade (CK) dataset [11] and its extended version (CK+) [12], the Oulu-CASIA dataset [13], the AffectNet dataset [14], the Acted Facial Expressions in the Wild (AFEW) dataset, and its static version: the Static Facial Expressions in the Wild (SFEW_2.0) dataset [15,16], and the Facial Expression Recognition 2013 (FER2013) [17]. Nevertheless, Facial Emotion Recognition (FER) has remained as an active research topic during the past decades due to various challenging factors such as illumination changes, head pose, head motion, movement blur, age, gender, and skin color [18]. In fact, FER is still difficult particularly in-the-wild as well as in unconstrained real-life environments. Early approaches for automatic facial expression recognition [19] usually perform quickly and accurately in indoor environments, but they frequently drop in performance under real-world conditions [20]. Therefore, there are still several challenging issues. Indeed, most of studies have based the hand-crafted feature extraction approaches completely on human experience, and that fact made them so complex in some real applications. Consequently, it is hard to extract prominent features using the classical methods. To deal with this challenge facing the quick progress of emotion recognition techniques, and in order to achieve higher accuracy, recent investigations are further motivated to develop FER systems based on deep learning techniques. Thus, investigating deep neural network models for facial expression analysis has become the hottest subject in recent facial recognition works [21]. In fact, feature learning allows deep networks to learn a broader range of facial features than earlier approaches, including rotation variation and illumination changes, and it has turned out that Convolutional Neural Networks (CNN) trained for facial expression recognition can learn facial features reflecting those suggested by the psychologist Ekman [22].

Overall, several recent works have effectively dealt with FER issues using CNN [23]. Nevertheless, CNN models elucidate several limitations deserving more attention such as the accuracy rate that could be higher, especially in-the-wild. To cope with this limitation, we mainly focused on features provided by different CNN models, and on the ability of each model to achieve high precision rates separately. Our concept aims to achieve the resourcefulness by having multiple resources, not from having only one intelligent. Subsequently, we propose in this work to build upon the fusion of deep features supplied by different CNN models. More precisely, we have studied the Resnet101, which ensured its efficiency in terms of learning with the depth of

the layers thanks to the use of residual learning networks. Moreover, the VGG19, which is a shallow model but with a remarkable amount of parameters, as well as the GoogleNet, which insures a balance between efficiency and speed of learning while reducing the parameters number of the network, are also investigated. In fact, the proposed method follows a standard FER scheme where face images are normalized, then augmented. Thereafter, the features from the pre-processed images are extracted using pre-trained CNN architectures and finally classified via an SVM classifier. The proposed method focuses on a layer-based feature selection from each pre-trained model separately. The concatenation includes the three feature vectors selected from different layers into a single final vector. The suggested scheme ensures the complementarity of facial expression features extracted from the three pre-trained architectures. This scheme is composed mainly of two phases: training and validation. During the training phase, images are pre-processed, then faces are detected and finally features are extracted from each model and then concatenated into a single vector to be fed to an SVM classifier for the training phase. The same pipeline is followed during the validation process. In fact, the main contributions of this work are twofold:

- We have applied three pre-trained neural networks in order to extract complementary features driven into multichannel solution with a personalized freezing weight during the training phase. A layer-based feature selection is performed from each pre-trained model separately. A layer search is performed from the last five layers including the FC ones. The layer that provides the best features is selected and the features it provides are retained.
- The final feature vector is formed by concatenating the features retained from the different pre-trained models. The concatenation phase has allowed to obtain a single model gathering the most relevant extracted facial information of the three basic models. The overall error rate is reduced compared to each single model since the failure percentage of one model could be fulfilled with that of another one.

Extensive experiments have been carried out on the most challenging FER datasets available today (JAFFE dataset of Japanese Female images, the Extensive Cohn-Kanade (CK+) dataset, the Facial Expression Recognition 2013 dataset (FER2013), and the SFEW_2.0 dataset of static images in the wild), and the proposed method has led to very promising results.

The remainder of this paper is organized as follows: Section 2 briefly reviews relevant existing FER methods. In Sect. 3, we describe the proposed method. In Sect. 4, an overview of datasets used in this work is outlined before providing experiments and performance comparison with

relevant state-of-the-art methods. Finally, conclusions and future research directions are given in Sect. 5.

2 Related work

A standard FER system involves essentially three key components, namely face detection and pre-processing, feature extraction, and classification. Face detection aims to determine the location and the size of the human face, or faces, within the input image [24]. The most widely used methods for face detection include MTCNN [25], Dlib [26], the eigen-face techniques [27], and the Viola-Jones algorithm [28]. Although face detection is an essential procedure enabling feature extraction, image pre-processing is usually required for the alignment and the normalization of the visual semantic information conveyed by the face. Its primary function is to ignore all variations irrelevant to facial expressions such as different backgrounds, illuminations, and head poses; fairly common in unconstrained scenarios; and to keep as much meaningful features as possible [29]. The second stage, which is feature extraction, intends to extract facial features from the pre-processed images of the detected faces [30]. The third stage is the classification of the extracted facial features into one of the basic emotion classes. Unlike the traditional methods where the feature extraction stage is independent of the feature classification one, deep networks can perform FER in an end-to-end manner [29]. Indeed, the way how facial changes are typically extracted into features [31] facilitates the emotion prediction for FER systems. In the remaining of this section, an overview of various FER works is presented briefly, while focusing on those that have been validated on the JAFFE, the CK+, the FER2013, and/or the SFEW_2.0 datasets. These works have been categorized, according to the adopted feature extraction approach, into three major groups: hand-crafted features, deep learning features and hybrid ones.

2.1 Hand-crafted features

First emotion recognition works have been based on hand-crafted feature representation methods, which are commonly divided into two categories: features based on templates (or appearance features) and geometric features. The appearance feature extraction methods (e.g. Gabor filter [32], Local Binary Pattern (LBP) [33], Histogram of Oriented Gradients (HOG) [34]...) are applied on the totality of the face image, whereas the geometric feature-based methods commonly exploit landmark points in order to calculate geometric distances between face regions [35]. It is worth noting that most of existing hand-crafted methods use a combination of these two approaches [36]. For instance, Zhang et al. [37] have cropped images of size 110×150 pixels after detecting

automatically the faces based on a set of rectangular Haar-like features. Then, features have been extracted using local binary patterns before applying the Local Fisher Discriminant Analysis (LFDA) in order to produce a representation of extracted data of low dimension. An accuracy of 90.7% has been reached by this method on the JAFFE dataset. Likewise, Abdulrahman and Eleyan [38] have focused their contribution on the feature extraction step. The conceived system has been based on LBP as feature extractor and the Principal Component Analysis (PCA) for the dimensionality reduction of the feature vectors. These vectors are then fed to a Support Vector Machine (SVM) for the classification. Experiments were carried out on the JAFFE and the MUFEE datasets and the method has proved to be efficient at 87% and 77%, respectively. Alshamsi et al. [39] have opted for the Hausdorff distance for the pre-processing and the face detection, followed by a combination of facial landmarks and centers of gravity for the feature extraction. Then, an SVM classifier has been applied while reaching an accuracy of 96.3% on the CK+ dataset, 91.9% on the JAFFE dataset, and 90.8% on the KDEF dataset. Differently, the FER system designed by Gite et al. [40] detects faces from facial images using the Viola-Jones algorithm. Then, a combination of geometric and appearance-based techniques has been explored in order to extract reliable features. In fact, the authors have investigated the coordinates of face landmarks before reducing the dimensionality of the feature vector using the principal component analysis. The method has been validated on the extended Cohn-Kanade (CK+) dataset and a recognition accuracy of 93%, using an SVM classifier, has been recorded. However, this FER system still struggled with the common issues of handling real-world conditions such as head movement, various lighting conditions, and low-intensity expressions. Overall, the major issues of the hand-crafted methods can be mainly summarized in the failure of low-level features to extract relevant local facial information, and the incapacity to capture high level salient information, notably under in-the-wild conditions such as different head positions, complex backgrounds, different distances from the camera, multi-face scenes, subject movement, and low lightness conditions.

2.2 Deep learning features

The swift progress of deep learning models has motivated researchers to introduce deep neural networks within the framework of FER systems. Therefore, in the last decades, most of works have leaned toward the use of deep learning techniques for FER purposes [41,42]. Indeed, a large proportion of the relevant FER systems have relied on CNNs because of their performance and flexibility [43]. In particular, CNN architectures have proved to be more robust, than the Multi-Layer Perceptron (MLP), to face location changes as well as to scale variations, especially in the case of pre-

viously unseen faces and pose variations [44]. In addition to CNN, Deep CNN (DCNN) [45], Deep Belief Networks (DBN) [46], Deep Auto-Encoder (DAE) [47], Recurrent Neural Networks (RNN) [48], Generative Adversarial Networks (GAN) [49], and recently transfer learning-based frameworks [50], have been successfully investigated for facial emotion recognition. For instance, Shaees et al. [51] have performed a quantitative comparison between an FER method that is fully based on transfer learning, using pre-trained CNN, with an hybrid FER method based on a mixture of deep learned features, which are extracted using transfer learning, along with mainstream classification. They chose the AlexNet pre-trained CNN architecture, for their first method. However, a multiclass SVM had been adopted as classifier for the second method. They evaluated their methods on two datasets, namely NVIE and CK+, and they achieved for the first method the recognition rates of 91.5% and 90.1%, respectively. For the second method, an increase till 99.3% (*resp.* 98.3%) on the NVIE (*resp.* the CK+) dataset has been achieved. In the same context of deep learning approaches, Zhang et al. [52] have proposed two FER methods, both are based on deep convolutional neural networks of double-channel weighted mixture (WMDCNN) structure. However, the first method is based on static images and the second one is based on image sequences while adding long short-term memory (WMCNN-LSTM). The facial regions in the designed systems are detected by the AdaBoost method, and thereafter cropped and rotated, and only faces are kept by masking the other areas. The experimental results of the WMDCNN network on the CK+, the JAFFE, the Oulu-CASIA and the MMI datasets have achieved average recognition rates of 98.5%, 92.3%, 86%, and 78.24%, respectively. Nevertheless, the WMCNN-LSTM architecture has achieved an average recognition rate of 97.5% on the CK+ dataset, of 88% on the Oulu-CASIA dataset and of 87.1% on the MMI dataset. Differently, Minaee et al. [9] have introduced a deep learning approach based on attentional convolutional networks while adding a visualization technique in order to specify the most expressive regions related to emotions in the faces' images. The proposed method has been evaluated on four datasets (FER-2013, Facial Expression Research Group (FERG), CK+ and JAFFE), and recognition rates of 70.02%, 99.3%, 98.0%, and 92.8%, respectively, have been reported. Chen et al. [53] have used a Deep Sparse Autoencoder Network (DSAN) for learning facial features, and a Softmax Regression (SR) for the classification of the facial expressions. An average emotion recognition of 94.761% has been reached by evaluating the method on the JAFFE dataset. Likewise, the FER system of Li et al. [31] has been conceived based on convolutional neural networks for feature extraction, preceded by a pre-processing phase including a new face cropping and rotation technique. The evaluation of this system has been performed on the CK+ and

the JAFFE datasets, and recognition accuracies of 97.38% and 97.18% have been recorded, respectively. However, deep learning methods typically require large numbers of training instances, what presents the transfer learning as an attractive approach for the in-the-wild FER.

2.3 Hybrid features

Although the success of automated FER systems based on deep learning architectures, many researchers have valued that the traditional extracted features (hand-crafted features) contain relevant information that capture texture, shape, and appearance information describing facial expressions. They consider that hand-crafted and deep learning features are complementary. Therefore, hand-crafted features can be effectively combined with deep learned features in order to further improve the robustness as well as the accuracy of FER, especially that hybrid methods are present in psychological mechanisms that recognize facial expressions [54]. For instance, a Deep Action Units Graph Network (DAUGN) has been investigated for facial expression recognition in [54]. The introduced network is based on a segmentation strategy that divides faces into action units, and CNN is thereafter used in order to fuse the local-appearance and global-geometry features. The proposed FER system has been evaluated on the CK+, the MMI, and the SFEW_2.0 datasets and has achieved 97.67%, 80.11% and 55.36%, respectively, as accuracy rates. The results obtained are competitive comparing to others works, but are still insufficient for in-the-wild facial images. Similarly, Fan and Tjahjadi [55] have proposed a hybrid framework based on deep features learned using convolutional neural networks, and hand-crafted features including shape and appearance descriptors. In fact, in order to collect the hand-crafted features while describing the local facial properties, shape descriptors from facial landmarks, related to the eyes, the nose, and the mouth, have been combined with PHOG features. The framework achieved an accuracy of 92.5% on the CK+ dataset. However, this framework has been validated on only one dataset putting in question its robustness as well as its overfitting risk. Sun and Lv [56] have also chose a hybrid model for facial expression recognition. They have combined Scale-Invariant Feature Transform (SIFT) descriptors with deep learning features extracted from a CNN model. The method has been validated on the CK+ dataset and has achieved an accuracy of 94.82%. The cross-dataset experiments on the JAFFE dataset have achieved an accuracy of 48.90%. Likewise, the FER method of Gogić et al. [57], called LBF-NN, has combined local binary features with deep learned features via a Gentle Boost Decision Trees Neural Network (GBDTNN). The extracted hand-crafted features have been based on facial landmarks detected from cropped facial images. The performance of the method has been eval-

Table 1 Summary of relevant studied works for FER in the JAFFE, the CK+, the SFEW_2.0, and/or the FER2013 datasets using hand-crafted, deep learning and hybrid features

Hand-crafted features					
Studies	Zhang et al. [37]	Abdulrahman and Eleyan [38]	Alshamsi et al. [39]	Gite et al. [40]	Dhall et al. [59]
Techniques	LBP, LFDA, SVM	PCA, LBP, SVM	Hausdorff Distance, facial Landmarks COG, SVM	Viola & Jones, facial landmarks, PCA, SVM	Local Binary Pattern-Three Orthogonal Planes (LBP-TOP), SVM
JAFFE	90.7%	87%	91.9%	-	-
CK+	-	-	96.3%	93%	-
SFEW_2.0	-	-	-	-	39.13%
FER2013	-	-	-	-	-
Deep learning features					
Studies	Shaees et al. [51]	Zhang et al. [52]	Minacee et al. [9]	Chen et al. [53]	Li et al. [31]
Techniques	AlexNet pre-trained, CNNs, Multiclass SVM	AdaBoost, WMCNN-LSTM	Attentional convolutional Network	Deep neural network	CNN, softmax regression
JAFFE	-	92.3%	92.8%	94.76%	97.18%
CK+	90.1%	98.5%	98%	-	97.38%
SFEW_2.0	-	-	-	-	-
FER2013	-	-	70.02%	-	-
Hybrid features					
Studies	Liu et al. [54]	Fan and Tjahjadi [55]	Gogic et al. [57]	Alreshidi et al. [58]	
Techniques	Action units, CNN	Facial Landmarks, PHOG, CNN	LBF Gentle Boost Decision Trees Neural Networks	AdaBoost	Neighborhood difference features
JAFFE	-	-	85.88%	-	-
CK+	97.67%	92.5%	96.48%	-	-
SFEW_2.0	55.36%	-	49.31%	57.7%	-
FER2013	-	-	-	-	-

uated on four datasets: CK+ with 96.48% of accuracy rate, 73.73% for MMI, 85.88% for JAFFE and an accuracy of 49.31% for SFEW_2.0. Nevertheless, the performance of the method is quite limited for the case of in-the-wild images, since facial expressions in nature are dynamic and change in intensity. Similarly, Alreshidi and Ullah [58] have conceived their facial emotion recognition system using hybrid features. They have extracted Neighborhood Difference Features (NDF) obtained from faces detected with AdaBoost cascade classifiers. They have tested the performance of their approach on the SFEW_2.0 and the RAF datasets, and they have achieved a precision rate of 57.7% for SFEW_2.0 and of 59.0% for RAF. Overall, in-the-wild facial expression recognition methods exclusively based on deep learned features have proved to be more effective than that methods combining such features with hand-crafted ones.

Table 1 encompasses some relevant research studies, ranging from early works up to more recent ones, for each category of features (hand-crafted, deep learning and hybrid methods). The selected works have been collected based on the datasets they used to validate their studies (JAFFE, CK+, SFEW_2.0 and/or FER2013). It is clear that the investigated hand-crafted features (e.g. LBP, PCA, LFDA. . .) have not given sufficiently descriptive patterns of facial expressions, whereas deep learning methods show a remarkable improvement of the precision rate, especially under in-the-wild contexts, up to 18.57%. However, the margin for improvement is still possible, especially in real condition environments. The contribution detailed in this work focuses on the transfer learning from recent deep learning architectures in order to introduce effective solutions for the implementation of FER systems. The most relevant deep face features are studied by challenging several deep architectures in the context of in-the-wild FER. In fact, the suggested method aims to fuse relevant features from several pre-trained CNN models in order to use them in a multichannel solution for the recognition of in-the-wild human facial expressions. To the best of our knowledge, it is the first time that deep learning features extracted from pre-trained architecture in the context of in-the-wild conditions are investigated and fused into a single solution to improve FER accuracy.

3 Proposed method

This section details the proposed method for FER in-the-wild. The method performs the FER task based on multichannel convolutional neural network, using dual deep learning networks. The first one is a DL as feature extractor based on transfer learning techniques. It uses three pre-trained CNN models namely VGG19 [60], GoogleNet [61], and ResNet101 [62]. The second one is a DL as a transformer. It consists to select the richest features' layer from each model.

The three resulting vectors are thereafter concatenated into a single vector representing the final feature vector to be fed to an SVM classifier in order to predict the emotion class of the input image. The proposed method aims to gather the most relevant features extracted from the VGG19, the GoogleNet, and the ResNet101 networks. It aims to exploit the complementarity of the extracted features from the three models in order to reduce the error rate. In what follows, we describe the different steps of the proposed emotion recognition procedure. In fact, the input images are pre-processed, before detecting the faces. After that, the three pre-trained CNN models are used for feature extraction. Then, the richest features from each model are selected and concatenated into a single vector representing the final feature vector, which is fed to the SVM classifier.

3.1 Pre-processing and data augmentation

For this study, the JAFFE, the CK+, the SFEW_2.0, and the FER2013 datasets have been investigated for the training and the evaluation. All the used datasets comprise face images with seven basic facial expressions (Anger, Surprise, Fear, Disgust, Happiness, Sadness, and Neutral). Dataset samples are shown in Fig. 1, whereas Fig. 2 illustrates in more details the proposed method steps through its instantiation for the JAFFE dataset.

In fact, input images are firstly converted into RGB space and then normalized by modifying the range of intensity values in order to ensure illumination change robustness [63]. Non-face parts and useless regions are thereafter removed from normalized images in order to keep only face regions. This pre-processing step is important to enhance the image recognition performance. In our case, the Viola & Jones face detection algorithm [28], which is known for its robustness especially in the case of frontal images, is used in order to localize the face regions and to crop them from the entire images composing the used datasets. Furthermore, since a convolutional neural network requires a large amount of data to reach better accuracy, the performance of the model could be improved by Data Augmentation (DA) solutions [64]. In fact, the more important number of samples the dataset contains, the more features can be extracted from them, and the more the model can be improved in performance. Thus, as account of the small size of some public FER datasets, DA techniques are commonly used to increase the sizes of the datasets. Mostly, translations, rotations and skewing DA techniques have shown their benefits while being computationally efficient [65]. In our case, the data augmentation step consists of creating new images from each cropped image, using the following augmentations: horizontal and vertical translations, horizontal reflection, and random image rotations with a rotation angle in $[-10^\circ, 10^\circ]$ (Fig. 3). It is worth noting that data augmentation was applied only on the JAFFE

Fig. 1 Prototypical facial expression images from the JAFFE dataset (first column), the CK+ dataset (second column), the SFEW_2.0 dataset (third column), and the FER2013 dataset (fourth column)

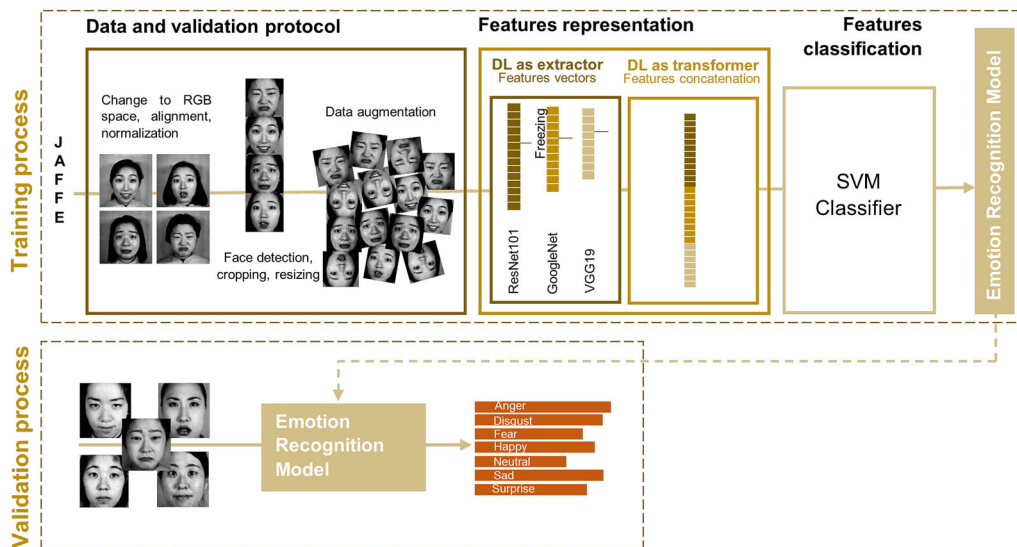


Fig. 2 Technical steps of the proposed FER method

and the SFEW_2.0 datasets, which include respectively 213 and 1230 images, because of their reduced numbers of samples compared to the CK+ and the FER2013 datasets, which include 5414 and 7178 images, respectively. Figure 3 illustrates some samples of the JAFFE and the SFEW_2.0 datasets, before and after applying the data augmentation.

3.2 Feature extraction

After resizing the input images in order to fit the input size of the pre-trained models, which is $224 \times 224 \times 3$, the feature extraction part of the proposed method is composed of two modules. The first one, called “DL as extractor”, consists on extracting features from the pre-processed facial images. To this end, a transfer learning has been applied while benefiting from the advantages of several relevant CNN models. The second module, called “DL as Transformer”, consists in concatenating the most relevant features selected from each single model to form the final prediction vector. The details of the two proposed modules are discussed in what follows.

- DL as extractor (CNN feature extraction): So as to represent the numerical information behind facial expressions, we have performed transfer learning on CNN models, which were pre-trained on 1000 classes from the ImageNet dataset, in order to discriminate between the seven emotional classes. In fact, we have tested several well-known deep learning models (ResNet50, ResNet101, VGG16, VGG19 and GoogleNet), which have already shown their effectiveness in several state-of-the-art FER works [9,66,67], on the challenging JAFFE dataset in order to assess their performance for the in-the-wild context. For more stable results, we have run the tested models twenty times. The mean and the standard deviation σ (1) have been calculated in order to choose the most appropriate models in terms of performance (*i.e.* highest accuracy means) as well as of stability (*i.e.* smaller standard deviations). For each studied model, the four best recognition rates, their mean and standard deviation are shown in Table 2. According to this Table, the RestNet101 has recorded the highest accuracy mean with the lowest standard deviation value, followed by the VGG19. The ResNet50

Fig. 3 Illustration of the different geometric DA techniques applied on the SFEW_2.0 (first row) and the JAFFE (second row) datasets



and the GoogleNet models have comparable mean value and standard deviation values. In this case, the choice of the third model was based on the mean of the three best recognition rates which gives the advantage to the GoogleNet model. For reasons related to the size of the final feature vector, with regards to the curse of dimensionality issue, and to have an odd number of sources, we opted for the choice of three models among the five tested ones for feature extraction. Thus, the realized experiments conducted us to choose the ResNet101, VGG19 and GoogleNet models in order to guarantee the most stable results in-the-wild context and therefore the most robust features.

$$\sigma = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2}, \quad (1)$$

where x_i denotes the recognition rates, μ is mean of the best recognition rates and n is the total number of experiences. Transfer learning techniques were then applied to these pre-trained CNN models while freezing weights at a personalized range of shallow layers, which does not capture relevant information. Freezing weight technique is applied for each model apart according to its depth, in order to keep only the relevant image properties for the training phase. This first step of the method freezes some shallow layers and keeps others deeper containing important data and having more ability to learn discriminant features. Freezing these layers aims to gain training time, and especially to eliminate less reliable features while retaining only relevant ones that perform more accurate recognition. The deep features extracted from the three models will be used afterward to train the SVM classifier. Furthermore, in order to confirm the suitability of the three chosen CNN models for the context of FER in gen-

eral, and not only for the in-the-wild context, we have also evaluated them separately on the CK+, the SFEW_2.0 and the FER2013 datasets. Each model has been tested on all three datasets, and the experiences were repeated twenty times while reporting the four best recognition rates (Table 3). We have also calculated the standard deviation and the mean of the recognition rates (Table 4). Indeed, for the JAFFE dataset, recognition rate reached 85.71% for the VGG19, 83.33% for the GoogleNet, and 85.71% for the ResNet101. For the CK+ dataset, recognition rates of 89.19% for the VGG19, 89.37% for the GoogleNet and 92.70% for the ResNet101 were recorded. For the SFEW_2.0 dataset, lower recognition rates were scored: 54.07% for the GoogleNet, 57.72% for the VGG19, and 60.57% for the ResNet101 models. Finally, for the FER2013 dataset, the VGG19 achieved an accuracy of 58.22%, 53.69% for the GoogleNet, and 55.57% for the ResNet101. Accuracies achieved using the three test CNN architectures are relatively good and promising for each one separately for the datasets taken in controlled environments but remain relatively low for uncontrolled environment (SFEW_2.0 and FER2013 datasets). However, after focusing on the confusion matrices of the three models on the SFEW_2.0 dataset, we have noticed that where one or two of the models fail, there is at least one that performs well. For example, the GoogleNet model fail to recognize the disgust emotion, whereas the ResNet101 model scores 28.6% for recognizing this emotion for the SFEW_2.0 dataset. Detailed results of the confusion matrices, which are illustrated later in the experimental result section, confirm this finding. This fact prompted us to investigate this complementarity while selecting the most suitable features from each model.

Table 2 Comparison results of five models applied on the JAFFE dataset

Models	Four best	Recognition	Rates (%)	Mean	Standard deviation (SD)
ResNet50	69.05	71.43	76.19	78.75	73.86
ResNet101	80.95	83.33	83.33	85.71	83.47
VGG16	64.29	61.90	66.67	76.19	67.26
VGG19	78.57	80.95	83.33	85.71	82.14
GoogleNet	76.19	73.81	71.43	83.33	76.19

Table 3 Four best emotion recognition rates of VGG19, GoogleNet and ResNet101 on the JAFFE, the CK+, the SFEW_2.0, and the FER2013 datasets (best values are in bold)

	VGG19			GoogleNet				ResNet101				
CK+	76.16	87.62	89.19	86.60	88.91	88.26	89.37	88.80	89.19	91.77	90.94	92.70
SFEW_2.0	57.32	57.14	56.91	57.72	48.37	51.22	51.63	54.07	57.45	53.25	56.91	60.57
FER2013	55.43	55.71	56.13	58.22	51.46	50.42	53.55	53.69	50.84	52.21	54.39	55.57

Table 4 The obtained recognition rates (mean and standard deviation (SD)) using the VGG19, the GoogleNet, and the ResNet101 models

Models	Datasets	Mean	Standard deviation (SD)
VGG19	JAFFE	82.14	2.66
	CK+	84.89	5.13
	SFEW_2.0	57.27	0.30
	FER2013	56.37	1.09
GoogleNet	JAFFE	76.19	4.46
	CK+	88.84	0.40
	SFEW_2.0	51.32	2.02
	FER2013	52.28	1.39
ResNet101	JAFFE	83.47	1.70
	CK+	91.15	1.30
	SFEW_2.0	57.04	2.60
	FER2013	53.25	1.84

- DL as Transformer (Feature concatenation): Several tests have been performed in order to choose, for each model, the most suitable layer for extracting the discriminant features. Firstly, the features have been extracted only from the Fully Connected (FC) layers. Afterward, the subsequent tests have shown that more discriminate features can be selected from other layers than the FC ones, notably the pooling layers, which preserve the most essential features of facial images. Thus, the layer-based feature selection process was focused on the five last layers of each model. The process has been empirically validated, and several tests have been carried out in order to select the most appropriate combination of feature layers for each of the three pre-trained models. Those layers contain quality features which help to increase the accuracy of the facial expression recognition model. The five best layers' combinations, in terms of recognition accuracy, from which the features were extracted, are summarized

in Table 5 for each of the four datasets. As illustrated in this table, the layers retained from the three models for the extraction of features depend on the dataset, which explains the difference in terms of the number of features retained for each dataset. For instance, the Drop7, the Fc7 and the pool5 layers, respectively, selected from the VGG19, the GoogleNet and the ResNet101 models, have been retained for feature concatenation for the case of the CK+ dataset. In fact, this is the best layer combination that gave an accuracy of 98.80%.

Nevertheless, the results illustrated in Table 5 show that the pooling layers contain more relevant features compared to the fully connected ones. In the majority of the cases, combining two pooling layers from two different models with a fully connected layer from the third model gave more efficiency than the combination of two fully connected layers with one pooling layer as well as than combining three fully connected layers. At the end of this stage, three vectors for each model (one for each dataset) corresponding to the highest recognition rates are retained. In fact, given the three feature vectors corresponding to the three pre-trained models, for each dataset, the concatenation module aims to construct, for each dataset, a single feature vector from the three sets of features retained from each CNN model. To perform that, we based in this study on the selection of the most significant layer for each model in order to extract the most relevant information for the emotion classification. The vectors extracted from each model are concatenated to form a single vector as shown in Fig. 4, where the number of extracted features for each dataset is also provided. Thus, once the layer from which the features is selected is chosen for each model, the concatenation is applied to form a final single feature vector that will be fed to the SVM classifier in order to predict the emotions of the test facial images. In fact, for the CK+ dataset, 6151 features have been retained from the three models (3079 features from the ResNet101 model,

2048 features from the GoogleNet model, and 1024 features from the VGG19 model), whereas 3079 features have been selected for the JAFFE dataset (1790 from the ResNet101 model, 521 from the GoogleNet model, and 768 from the VGG19 model). However, 3328 features have been kept for the SFEW_2.0 dataset (2048 features from the ResNet101 model, 256 from the GoogleNet model, and 1024 from the VGG19 model). For the FER2013 dataset, 10787 features have been retrained (6144 from the ResNet101 model, 2048 from GoogleNet, and 2595 from VGG19).

3.3 Emotion's classification

After forming the final vector resulting from the concatenation of the features selected from the three initial vectors, the classification step consists to associate each studied image to the corresponding emotion class. As mentioned previously, the test images are different from the training images and the number of samples is smaller. Instead of the classification layers of the models, a linear support vector machine has been used as a classifier of emotions. In the case of few samples per class, the SVM shows its efficiency to classify into different classes all new instances derived from the test set based on the emotions learnt. Due to the relevance of the data obtained in the extraction and the concatenation steps, we do not need to adopt a kernel for the transformation of features. Thus, SVM is used to find the optimal hyperplane that maximizes the distance between it and the closest data point called the margin of separation. In fact, as we are faced with a multiple classification problem (non-binary), we used in this work linear SVM, while following the one-vs-rest strategy that implements the multiclass SVM.

4 Experimental results

Having a high number of labeled data is a necessity to train a neural network in order to enable it to handle the curse of dimensionality problem [68]. In this work, four publicly available datasets have been used. In fact, the investigated datasets are as follows: (i) the Extended Cohn-Kanade dataset (CK+), which is conceived in laboratory-controlled conditions, contains mixture of posed and spontaneous emotions, (ii) the Japanese Female Facial Expression dataset (JAFFE), also conceived in laboratory-controlled conditions, contains only posed emotions, (iii) the Static Facial Expression in the Wild dataset (SFEW_2.0), and (iv) the Facial Expression Recognition 2013 (FER2013), which illustrate spontaneous emotions taken under in-the-wild conditions. In what follows, we give a brief overview of these datasets before diving into the results.

1- Extended Cohn-Kanade dataset (CK+): This dataset is an extended version of the "CK" collection, which has been

released since 2000 in order to promote research works in the field of facial expression detection [11]. All images have been designed in controlled environments. The subjects are both male and female where 31% are men and 69% are women with their age range from 20 to 45 years [69]. The dataset includes 593 sequences of images varying in duration from 10 to 60 frames collected from 123 subjects. Every image has 640×490 or 640×480 pixels resolution and their total-ity express seven emotion categories: the six basic emotions (Anger, Disgust, Fear, Happiness, Sadness, Surprise) and one contempt [12].

2- Japanese Female Facial Expression dataset (JAFFE): It is a laboratory-controlled dataset. As a benchmark collection, the JAFFE dataset is composed of 213 grayscale facial expression images of 10 Japanese women. The dataset is categorized for seven expressions: Neutral plus the six basic emotional expressions (Anger, Disgust, Fear, Happiness, Sadness and Surprise). Each image size is 256×256 pixels, and each of the images is rated based on six emotion adjectives using 60 Japanese subjects; each expressor has 2–4 samples for each expression. In this dataset, the same expression of one person may differ greatly in different samples and distinct expressions may not be very distinguishable [70].

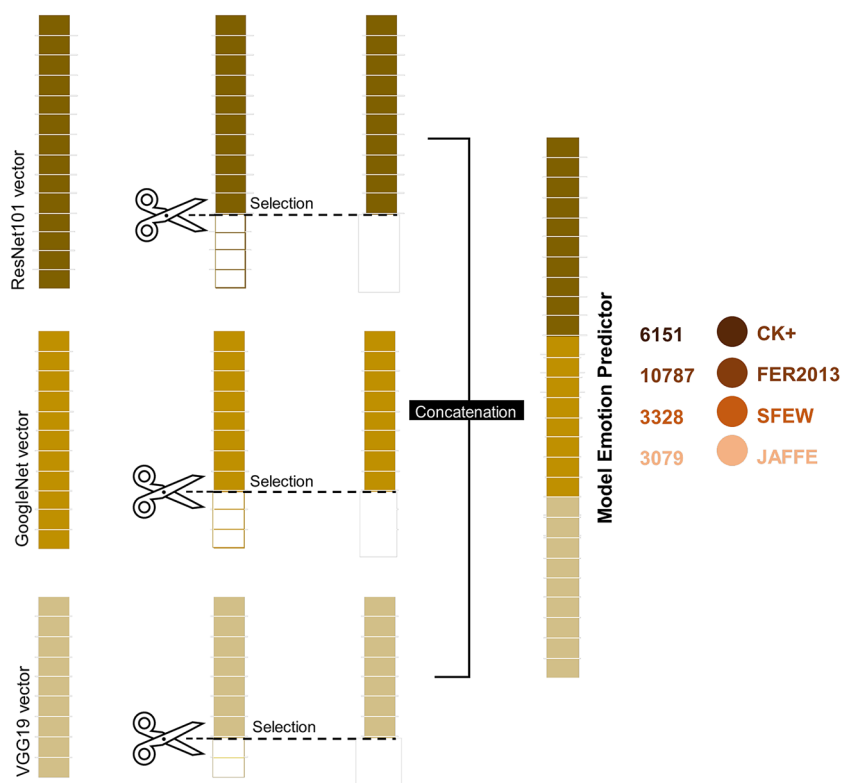
3- The Static Facial Expressions in the Wild (SFEW_2.0): It is a static dataset covering unconstrained facial expressions, different head poses, wide age range, varied face resolutions and focus making it close to real-world illumination. It has been extracted from the temporal dataset Acted Facial Expressions in the Wild (AFEW) and was firstly published in 2011 by Dhall et al. [71]. Consequently, it is analogous to the AFEW set except for its composition of static frames of the movies. In fact, each studied frame has been associated with an expression label (Angry, Disgust, Fear, Happy, Sad, Surprise, or Neutral) under close to real-world conditions. The SFEW_2.0 dataset contains 1766 images partitioned into 958, 436, and 372 images, for the training, the validation, and the test sets, respectively.

4- The Facial Expression Recognition 2013 (FER2013): This dataset has been developed by collecting face images available on the Internet, using the Google Image Search API. All images in this dataset have been captured in uncontrolled environments which made it a challenging standard benchmark within the framework of in-the-wild FER [67]. It contains 35,887 images belonging to the main seven emotions classes (4953 images for "Anger", 547 "Disgust" images, 5121 "Fear" images, 8989 "Happiness" images, 6077 "Sadness" images, 4002 "Surprise" images, and 6198 images for "Neutral"), while being divided into two sets: the training set and the test set [17]. However, the images are in gray scale with size restricted to 48×48 pixels.

Table 5 Top five layers' combinations for the four investigated datasets

Models	JAFFE			CK+			SFEW_2.0			FER2013		
	Layers	Final accuracy (%)	Layers	Final accuracy (%)	Layers	Final accuracy (%)	Layers	Final accuracy (%)	Layers	Final accuracy (%)		
VGG19, GoogleNet, ResNet101	Fc_1 Pool5-7x7_s1 Pool5	97.62	Drop7 Fc7 Pool5	98.80	Fc_1 Pool5-7x7_s1 Pool5	88.21	Fc7 Pool5-7x7_s1 Pool5	94.01	Fc7 Pool5-7x7_s1 Pool5	94.01		
VGG19, GoogleNet, ResNet101	Pool5 Fc7 Fc7	95.24	Fc7 Pool5-7x7_s1 Fc7	97.69	Fc7 Fc7 Pool5	84.60	Pool5 Fc7 Fc7	93.52	Pool5 Fc7 Fc7	93.52		
VGG19, GoogleNet, ResNet101	Fc7 Pool5-7x7_s1 Pool5	92.86	Fc7 Fc7 Pool	97.41	Fc7 Pool5-drop_7x7_s1 Pool5	79.67	Fc7 Pool5-drop_7x7_s1 Pool5	92.48	Fc7 Pool5-drop_7x7_s1 Pool5	92.48		
VGG19, GoogleNet, ResNet101	Fc7 Pool5 Fc7	90.48	Fc7 Pool5-7x7_s1 Pool5	96.93	Fc7 Fc7 Fc7	78.46	Fc7 Fc7 Fc7	91.50	Fc7 Fc7 Fc7	91.50		
VGG19, GoogleNet, ResNet101	Fc_2 Fc7 Pool5	90.02	Fc7 Fc7 Fc7	96.40	Fc_1 Pool5 Pool5	77.54	Fc_1 Fc7 Fc7	89.7	Fc_1 Fc7 Fc7	89.7		

Fig. 4 Principal of layers' selection and concatenation



4.1 Data preparation and validation protocol

For this study, the four datasets, JAFFE, CK+, SFEW_2.0, and FER2013 including, respectively, 213, 5414, 1230, and 7178 images, have been investigated. The images of the CK+ dataset have been manually divided into six classes of emotions, and the seventh class, which is “Neutral”, has been designed by collecting the first three sequences of emotion from each person of the six classes. We selected 5414 images from the five categories of emotions: happiness, fear, sadness, surprise, anger, disgust, while ignoring the class “contempt”. The JAFFE dataset have been also manually divided into seven classes of emotions, whereas the selected images from the SFEW_2.0 and the FER2013 datasets have been used as downloaded. Datasets are randomly split into training and testing samples with a split ratio of 80:20. Table 6 presents the numbers of samples for the training and the testing partitions, and the total number of images used for each dataset. All the CNN models have been trained for maximum 55 epochs. The ADAM optimizer has been applied for the GoogleNet and the ResNet101 models, while the SIGMOID has been used to optimize the VGG19 model. The initial learning rate was fixed as $1.e^{-4}$ for all the models.

The performance of the proposed method is presented on the above datasets. In fact, the produced results by the proposed multichannel CNN solution for facial emotion recognition are herein presented in two separate parts. The

first part of the results is related to the first feature extraction deep learning network. The second part gives the results of the final accuracy rates after selecting and concatenating features. It is worth mentioning that all accuracies are referring to testing accuracy on samples that are not included in the training. The outputs of the first deep learning network as extractor (first step), where for each model a freezing weight has been applied to certain blocks of layers during the training phase, are presented first. The confusion matrices summarize the prediction results for each emotion apart. They have been generated to assess and to unravel apart each pre-trained model. These matrices have been presented in this work for each pre-trained model apart and for the proposed model to firstly demonstrate that the used models are complementary and do not err in the same emotions and then to show that the feature concatenation can enhance the recognition rate for emotions that are hard to capture.

4.2 Results of the three models on the JAFFE dataset

Three feature vectors have been selected for the JAFFE dataset with an accuracy of 85.71% from VGG19, 83.33% from GoogleNet, and 85.71% from ResNet101. We report in Tables 7, 8 and 9 the corresponding confusion matrices, which show that the VGG19 model achieves 100% of recognition rate for four emotions (Fear, Happiness, Neutral and Surprise), whereas Anger emotion is recognized only with

Table 6 Numbers of samples for the four datasets

Datasets	Training samples	Testing samples	Total number of images
JAFFE	170	43	213
CK+	4331	1083	5414
SFEW_2.0	984	246	1230
FER2013	5742	1436	7178

Table 7 The confusion matrix of the VGG19 model on the test set of JAFFE

Output Class	Anger	Disgust	Fear	Happiness	Neutral	Sadness	Surprise	Accuracy
Anger	3 7.1%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	100% 0.0%
Disgust	0 0.0%	4 9.5%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	100% 0.0%
Fear	1 2.4%	1 2.4%	6 14.3%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	75.0% 25.0%
Happiness	0 0.0%	1 2.4%	0 0.0%	6 14.3%	0 0.0%	0 0.0%	0 0.0%	85.7% 14.3%
Neutral	0 0.0%	0 0.0%	0 0.0%	0 0.0%	6 14.3%	1 2.4%	0 0.0%	85.7% 14.3%
Sadness	2 4.8%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	5 11.9%	0 0.0%	71.4% 28.6%
Surprise	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	6 14.3%	100% 0.0%
	50.0% 50.0%	66.7% 33.3%	100% 0.0%	100% 0.0%	100% 0.0%	83.3% 16.7%	100% 0.0%	85.7% 14.3%
Target Class	Anger	Disgust	Fear	Happiness	Neutral	Sadness	Surprise	

Table 8 The confusion matrix of the GoogleNet model on the test set of JAFFE

Output Class	Anger	Disgust	Fear	Happiness	Neutral	Sadness	Surprise	Accuracy
Anger	6 14.3%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	1 2.4%	0 0.0%	85.7% 14.3%
Disgust	0 0.0%	6 14.3%	0 0.0%	0 0.0%	0 0.0%	1 2.4%	0 0.0%	85.7% 14.3%
Fear	0 0.0%	0 0.0%	6 14.3%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	100% 0.0%
Happiness	0 0.0%	0 0.0%	0 0.0%	5 11.9%	0 0.0%	0 0.0%	0 0.0%	100% 0.0%
Neutral	0 0.0%	0 0.0%	0 0.0%	1 2.4%	3 7.1%	0 0.0%	1 2.4%	60.0% 40.0%
Sadness	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	4 9.5%	0 0.0%	100% 0.0%
Surprise	0 0.0%	0 0.0%	0 0.0%	0 0.0%	3 7.1%	0 0.0%	5 11.9%	62.5% 37.5%
	100% 0.0%	100% 0.0%	100% 0.0%	83.3% 16.7%	50.0% 33.3%	66.7% 33.3%	83.3% 16.7%	83.3% 16.7%
Target Class	Anger	Disgust	Fear	Happiness	Neutral	Sadness	Surprise	

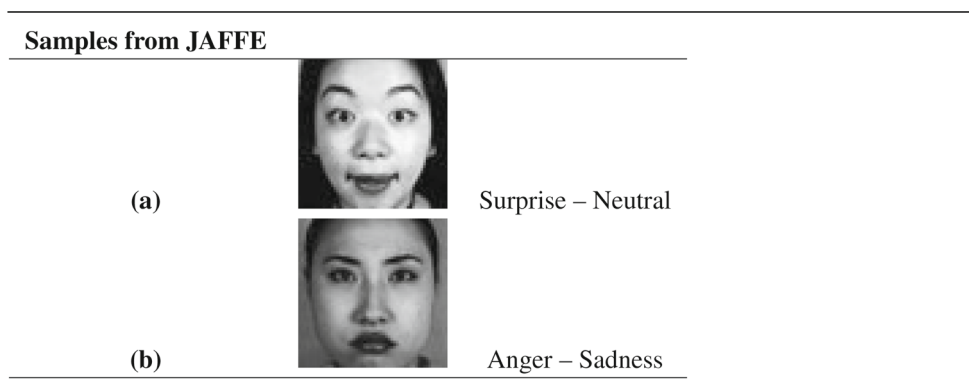
a rate of 50%. Disgust and Sadness have recognition rates of 66.7% and 83.3% respectively. However, the GoogleNet model achieves 100% of recognition rate for Anger and Disgust emotions, which are recognized only at 50% and 66.7% respectively by VGG19. GoogleNet also achieves 100% of recognition rate for Fear emotion.

The ResNet101 model recognizes 100% for Fear, Happiness, and Surprise emotions. It reaches 83.3% for the Disgust and Neutral emotions, and 66.7% of recognition rate for Anger and Sadness. Although GoogleNet does not reach a high accuracy for Happiness, Surprise, and Neutral emotions, ResNet101 has recognized 100% for Happiness and Surprise and has reached a rate of 33.3% for the case of the Neutral emotion. While the neutral class had an average recognition rate of 50% by GoogleNet, it reached an accuracy of 100% by VGG19 and a 16.6% better success rate for Sadness compared to GoogleNet and ResNet101. Overall, the recorded results show a complementarity between the three models recognizing the seven emotional classes. That fact allows us to conclude that some models classify correctly some emotions while other models misclassify the same emotions. This finding is illustrated by Table 10(a) which shows an image misclassified by the GoogleNet model and correctly clas-

Table 9 The confusion matrix of the ResNet101 model on the test set of JAFFE

Output Class	Anger	Disgust	Fear	Happiness	Neutral	Sadness	Surprise	Accuracy
Anger	4 9.5%	1 2.4%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	80.0% 20.0%
Disgust	0 0.0%	5 11.9%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	100% 0.0%
Fear	0 0.0%	0 0.0%	6 14.3%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	100% 0.0%
Happiness	0 0.0%	0 0.0%	0 0.0%	6 14.3%	1 2.4%	1 2.4%	0 0.0%	75.0% 25.0%
Neutral	0 0.0%	0 0.0%	0 0.0%	0 0.0%	5 11.9%	1 2.4%	0 0.0%	83.3% 16.7%
Sadness	2 4.8%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	4 9.5%	0 0.0%	66.7% 33.3%
Surprise	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	6 14.3%	100% 0.0%
	66.7% 33.3%	83.3% 16.7%	100% 0.0%	100% 0.0%	83.3% 16.7%	66.7% 33.3%	100% 0.0%	85.7% 14.3%
Target Class	Anger	Disgust	Fear	Happiness	Neutral	Sadness	Surprise	

Table 10 Misclassified JAFFE images (Original Class–Predicted Class)



sified by the VGG19 model. Table 10(b) presents also an example of image misclassified by the ResNet101 model and correctly classified by the GoogleNet model.

4.3 Results of the three models on the CK+ dataset

Tables 11, 12 and 13 gather the confusion matrices representing the accuracies of the resulting feature vectors of each studied model on the CK+ dataset. In fact, the global recognition rate is 89.19% for VGG19, 89.37% from GoogleNet, and 92.70 % from ResNet101. We have assessed the recognition rate by comparing the confusion matrices of the three pre-trained models, and it is clear that the VGG19 recognizes better the emotion “Happiness” with a recognition rate of 95.3% compared to GoogleNet and ResNet101 models. While for the “Fear” emotion, VGG19 and ResNet101 had the same recognition rate (=92.5%). The GoogleNet model recognizes better the emotion “Sadness” with an accuracy of 97.8%. For “Disgust” emotion, GoogleNet and ResNet101 achieved a recognition rate of 91.1%, while the Anger, Neutral and Surprise emotions have been recognized better with ResNet101 with accuracies of 97.7%, 85.8% and 95.6%, respectively. Similarly to the case of the JAFFE dataset, some images of this dataset are misclassified by one model but are correctly classified by another one. Table 14 shows some examples: images (a,b) are misclassified by GoogleNet but are correctly classified by ResNet101, whereas image (c) is correctly classified by ResNet101 and misclassified by VGG19, and image (d) is misclassified by GoogleNet but it is correctly classified by VGG19, and the final image (e) illustrates an example that is incorrectly classified by ResNet101 while being correctly classified by GoogleNet.

4.4 Results of the three models on the SFEW_2.0 dataset

Tables 15, 16 and 17 show the confusion matrices illustrating the emotion recognition rates of each model for the facial images of the SFEW_2.0 dataset, which are taken in

Table 11 The confusion matrix of the VGG19 model on the test set of CK+

		Confusion Matrix							
		Angry	Disgust	Fear	Happy	Neutral	Sad	Surprised	
Output Class	Angry	164 15.2%	0 0.0%	0 0.0%	0 0.0%	12 1.1%	0 0.0%	0 0.0%	93.2% 6.8%
	Disgust	0 0.0%	121 11.2%	0 0.0%	0 0.0%	12 1.1%	0 0.0%	1 0.1%	90.3% 9.7%
	Fear	0 0.0%	0 0.0%	86 7.9%	0 0.0%	9 0.8%	0 0.0%	0 0.0%	90.5% 9.5%
	Happy	0 0.0%	0 0.0%	0 0.0%	202 18.7%	15 1.4%	0 0.0%	1 0.1%	92.7% 7.3%
	Neutral	7 0.6%	14 1.3%	5 0.5%	10 0.9%	116 10.7%	4 0.4%	13 1.2%	68.6% 31.4%
	Sad	0 0.0%	0 0.0%	0 0.0%	0 0.0%	7 0.6%	85 7.9%	0 0.0%	92.4% 7.6%
	Surprised	0 0.0%	0 0.0%	2 0.2%	0 0.0%	5 0.5%	0 0.0%	191 17.7%	96.5% 3.5%
			95.9% 4.1%	89.6% 10.4%	92.5% 7.5%	95.3% 4.7%	65.9% 34.1%	95.5% 4.5%	92.7% 7.3%
		Target Class							

real conditions. In fact, the mean accuracies are as follows: 57.72% from VGG19, 54.07 % from GoogleNet, and 60.60 % from ResNet101. The VGG19 model achieves the best emotion recognition rate for Happiness and Surprise compared to GoogleNet and ResNet101, at 89.4% and 60.7% of accuracy, respectively. The GoogleNet model could not recognize the Disgust emotion; however, it was able to achieve the best recognition rates of 66.7% for the Sadness emotion, and 54.2% for the Fear emotion. The three emotions Anger, Disgust, and Neutral have been recognized better using the ResNet101 model, with the following rates 63.8%, 28.6% and 56.8%, respectively.

Table 18 illustrates some examples of images that are simultaneously misclassified by one model and correctly classified by another one. For instance, the image shown in Table 18(a) has been misclassified by ResNet101 but it was correctly classified by GoogleNet. The image in

Table 12 The confusion matrix of the GoogleNet model on the test set of CK+

		Confusion Matrix							
Output Class	Angry	157 14.5%	2 0.2%	1 0.1%	1 0.1%	6 0.6%	0 0.0%	0 0.0%	94.0% 6.0%
	Disgust	1 0.1%	123 11.4%	0 0.0%	1 0.1%	13 1.2%	0 0.0%	0 0.0%	89.1% 10.9%
	Fear	0 0.0%	0 0.0%	84 7.8%	0 0.0%	2 0.2%	0 0.0%	2 0.2%	95.5% 4.5%
	Happy	1 0.1%	0 0.0%	1 0.1%	194 17.9%	6 0.6%	0 0.0%	2 0.2%	95.1% 4.9%
	Neutral	10 0.9%	9 0.8%	7 0.6%	16 1.5%	138 12.8%	2 0.2%	18 1.7%	69.0% 31.0%
	Sad	2 0.2%	1 0.1%	0 0.0%	0 0.0%	8 0.7%	87 8.0%	0 0.0%	88.8% 11.2%
	Surprised	0 0.0%	0 0.0%	0 0.0%	0 0.0%	3 0.3%	0 0.0%	184 17.0%	98.4% 1.6%
			91.8% 8.2%	91.1% 8.9%	90.3% 9.7%	91.5% 8.5%	78.4% 21.6%	97.8% 2.2%	89.3% 10.7%
		Target Class							

Table 13 The confusion matrix of the ResNet101 model on the test set of CK+

		Confusion Matrix							
Output Class	Angry	167 15.4%	0 0.0%	1 0.1%	0 0.0%	2 0.2%	0 0.0%	0 0.0%	98.2% 1.8%
	Disgust	0 0.0%	123 11.4%	0 0.0%	0 0.0%	4 0.4%	0 0.0%	0 0.0%	96.9% 3.1%
	Fear	0 0.0%	0 0.0%	86 7.9%	0 0.0%	1 0.1%	0 0.0%	0 0.0%	98.9% 1.1%
	Happy	0 0.0%	0 0.0%	0 0.0%	196 18.1%	5 0.5%	0 0.0%	1 0.1%	97.0% 3.0%
	Neutral	4 0.4%	12 1.1%	4 0.4%	16 1.5%	151 14.0%	6 0.6%	8 0.7%	75.1% 24.9%
	Sad	0 0.0%	0 0.0%	0 0.0%	0 0.0%	5 0.5%	83 7.7%	0 0.0%	94.3% 5.7%
	Surprised	0 0.0%	0 0.0%	2 0.2%	0 0.0%	8 0.7%	0 0.0%	197 18.2%	95.2% 4.8%
			97.7% 2.3%	91.1% 8.9%	92.5% 7.5%	92.5% 7.5%	85.8% 14.2%	93.3% 6.7%	95.6% 4.4%
		Target Class							

Table 18(b) is misclassified by VGG19 and correctly classified by ResNet101. The last example in Table 18(c) is misclassified by the GoogleNet model and correctly classified by the VGG19 model.


4.5 Results of the three models on the FER2013 dataset

The three vectors selected for the FER2013 dataset in the first step have the following recognition rates: 58.22% from VGG19, 53.69% from GoogleNet, and 55.57% from ResNet101. The confusion matrices of the accuracies have been reported in Tables 19, 20 and 21. The best emotion recognition rate has been achieved for the Happiness emotion, at 81.70%, by the VGG19 model. The lowest recog-


nition rate is for the “Disgust” class with an accuracy of 18.20% achieved by the GoogleNet model. The ResNet101 and the VGG19 models achieved the same accuracy of 43.80% for the Anger class; however, GoogleNet recognized better the emotion. The emotions Fear and Sad are better recognized by the VGG19 model with accuracies of 37.10% and 55.40%, respectively, while the Surprise emotion reached 74.70% by the VGG19 model. The principal of complementarity is also confirmed with the FER2013 dataset results. Table 22 shows some examples of FER2013 images that are misclassified by one model and correctly classified by another model. In Table 22(a), the image is misclassified by the ResNet101 model, but it is correctly classified by GoogleNet, whereas the image in Table 22(b) is correctly classified by GoogleNet and misclassified by VGG19.

Table 14 Misclassified CK+ images (Original Class–Predicted Class)


Samples from CK+




(a) Anger – Neutral




(b) Disgust – Anger



(c) Surprise – Neutral



(d) Sad – Neutral



(e) Happy – Neutral

Table 15 The confusion matrix of the VGG19 model on the test set of SFEW_2.0

		Confusion Matrix								
Output Class	Angry	25 10.2%	1 0.4%	0 0.0%	1 0.4%	2 0.8%	1 0.4%	3 1.2%	75.8%	24.2%
	Disgust	0 0.0%	3 1.2%	0 0.0%	0 0.0%	2 0.8%	2 0.8%	0 0.0%	42.9%	57.1%
	Fear	2 0.8%	2 0.8%	11 4.5%	0 0.0%	2 0.8%	0 0.0%	1 0.4%	61.1%	38.9%
	Happy	10 4.1%	2 0.8%	1 0.4%	42 17.1%	9 3.7%	10 4.1%	0 0.0%	56.8%	43.2%
	Neutral	0 0.0%	2 0.8%	2 0.8%	0 0.0%	17 6.9%	0 0.0%	3 1.2%	70.8%	29.2%
	Sad	6 2.4%	2 0.8%	1 0.4%	2 0.8%	8 3.3%	27 11.0%	4 1.6%	54.0%	46.0%
	Surprise	4 1.6%	2 0.8%	9 3.7%	2 0.8%	4 1.6%	2 0.8%	17 6.9%	42.5%	57.5%
			53.2%	21.4%	45.8%	89.4%	38.6%	64.3%	60.7%	57.7%
		46.8%	78.6%	54.2%	10.6%	61.4%	35.7%	39.3%	42.3%	
	Target Class	Angry	Disgust	Fear	Happy	Neutral	Sad	Surprise		

Table 17 The confusion matrix of the ResNet101 model on the test set of SFEW_2.0

		Confusion Matrix								
Output Class	Angry	30 12.2%	3 1.2%	1 0.4%	1 0.4%	1 0.4%	2 0.8%	2 0.8%	75.0%	25.0%
	Disgust	1 0.4%	4 1.6%	0 0.0%	1 0.4%	2 0.8%	3 1.2%	1 0.4%	33.3%	66.7%
	Fear	0 0.0%	0 0.0%	13 5.3%	0 0.0%	0 0.0%	0 0.0%	3 1.2%	81.3%	18.8%
	Happy	3 1.2%	1 0.4%	5 2.0%	38 15.4%	7 2.8%	5 2.0%	2 0.8%	62.3%	37.7%
	Neutral	6 2.4%	2 0.8%	2 0.8%	2 0.8%	25 10.2%	6 2.4%	4 1.6%	53.2%	46.8%
	Sad	5 2.0%	3 1.2%	0 0.0%	3 1.2%	6 2.4%	25 10.2%	2 0.8%	56.8%	43.2%
	Surprise	2 0.8%	1 0.4%	3 1.2%	2 0.8%	3 1.2%	1 0.4%	14 5.7%	53.8%	46.2%
			63.8%	28.6%	54.2%	80.9%	56.8%	59.5%	50.0%	60.6%
		36.2%	71.4%	45.8%	19.1%	43.2%	40.5%	50.0%	39.4%	
	Target Class	Angry	Disgust	Fear	Happy	Neutral	Sad	Surprise		

Table 16 The confusion matrix of the GoogleNet model on the test set of SFEW_2.0

		Confusion Matrix								
Output Class	Angry	28 11.4%	2 0.8%	4 1.6%	3 1.2%	5 2.0%	1 0.4%	5 2.0%	58.3%	41.7%
	Disgust	2 0.8%	0 0.0%	3 1.2%	1 0.4%	3 1.2%	2 0.8%	2 0.8%	0.0%	100.0%
	Fear	3 1.2%	2 0.8%	13 5.3%	2 0.8%	4 1.6%	3 1.2%	5 2.0%	40.6%	59.4%
	Happy	4 1.6%	5 2.0%	0 0.0%	34 13.8%	2 0.8%	3 1.2%	1 0.4%	69.4%	30.6%
	Neutral	6 2.4%	2 0.8%	1 0.4%	5 2.0%	23 9.3%	5 2.0%	5 2.0%	48.9%	51.1%
	Sad	3 1.2%	2 0.8%	2 0.8%	2 0.8%	6 2.4%	28 11.4%	3 1.2%	60.9%	39.1%
	Surprise	1 0.4%	1 0.4%	1 0.4%	0 0.0%	1 0.4%	0 0.0%	7 2.8%	63.6%	36.4%
			59.6%	0.0%	54.2%	72.3%	52.3%	66.7%	25.0%	54.1%
		40.4%	100.0%	45.8%	27.7%	47.7%	33.3%	75.0%	45.9%	
	Target Class	Angry	Disgust	Fear	Happy	Neutral	Sad	Surprise		

Table 18 Misclassified SFEW_2.0 images (Original Class–Predicted Class)

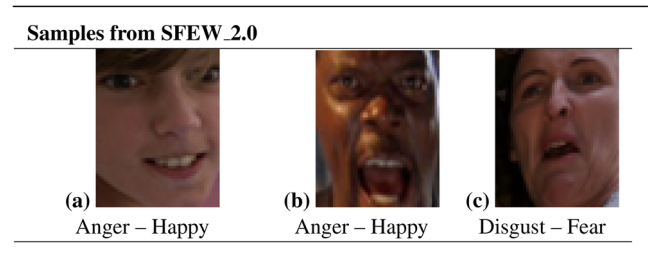


Table 19 The confusion matrix of the VGG19 model on the test set of FER2013

		Confusion Matrix								
Output Class	Angry	84 5.8%	9 0.6%	16 1.1%	8 0.6%	17 1.2%	28 1.9%	5 0.3%	50.3%	49.7%
	Disgust	2 0.1%	6 0.4%	1 0.1%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	66.7%	33.3%
	Fear	22 1.5%	1 0.1%	76 5.3%	12 0.8%	23 1.6%	30 2.1%	30 2.1%	39.2%	60.8%
	Happy	12 0.8%	3 0.2%	10 0.7%	290 20.2%	21 1.5%	11 0.8%	12 0.8%	80.8%	19.2%
	Neutral	18 1.3%	2 0.1%	28 1.9%	21 1.5%	133 9.3%	37 2.6%	5 0.3%	54.5%	45.5%
	Sad	45 3.1%	1 0.1%	54 3.8%	19 1.3%	51 3.6%	138 9.6%	5 0.3%	44.1%	55.9%
	Surprised	9 0.6%	0 0.0%	20 1.4%	5 0.3%	2 0.1%	5 0.3%	109 7.6%	72.7%	27.3%
			43.8%	27.3%	37.1%	81.7%	53.8%	55.4%	65.7%	58.2%
		56.3%	72.7%	62.9%	18.3%	46.2%	44.6%	34.3%	41.8%	
	Target Class	Angry	Disgust	Fear	Happy	Neutral	Sad	Surprised		

Another image misclassified by the ResNet101 while being correctly classified by the VGG19 model, is displayed in Table 22(c).

4.6 Results of the proposed model after feature extraction and concatenation on the four used datasets

The second key step of the proposed emotion recognition system in this study is the selection of features from each

Table 20 The confusion matrix of the GoogleNet model on the test set of FER2013

		Confusion Matrix							
Output Class	Angry	101 7.0%	6 0.4%	23 1.6%	14 1.0%	20 1.4%	28 1.9%	13 0.9%	49.3% 50.7%
	Disgust	3 0.2%	4 0.3%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	57.1% 42.9%
	Fear	21 1.5%	4 0.3%	72 5.0%	13 0.9%	29 2.0%	38 2.6%	29 2.0%	35.0% 65.0%
	Happy	16 1.1%	5 0.3%	14 1.0%	258 18.0%	18 1.3%	19 1.3%	12 0.8%	75.4% 24.6%
	Neutral	20 1.4%	2 0.1%	31 2.2%	35 2.4%	127 8.8%	44 3.1%	13 0.9%	46.7% 53.3%
	Sad	27 1.9%	1 0.1%	49 3.4%	31 2.2%	44 3.1%	116 8.1%	6 0.4%	42.3% 57.7%
	Surprised	4 0.3%	0 0.0%	16 1.1%	4 0.3%	9 0.6%	4 0.3%	93 6.5%	71.5% 28.5%
			52.6% 47.4%	18.2% 81.8%	35.1% 64.9%	72.7% 27.3%	51.4% 48.6%	46.6% 53.4%	56.0% 44.0%
		Target Class							

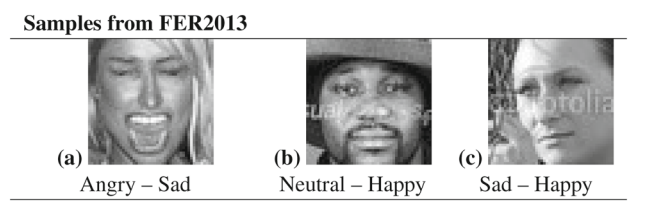
Table 23 Confusion matrix of the proposed method on the JAFFE dataset

		Confusion Matrix							
Output Class	Anger	6 14.3%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	100% 0.0%
	Disgust	0 0.0%	6 14.3%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	100% 0.0%
	Fear	0 0.0%	0 0.0%	6 14.3%	0 0.0%	1 2.4%	0 0.0%	0 0.0%	85.7% 14.3%
	Happiness	0 0.0%	0 0.0%	0 0.0%	6 14.3%	0 0.0%	0 0.0%	0 0.0%	100% 0.0%
	Neutral	0 0.0%	0 0.0%	0 0.0%	0 0.0%	5 11.9%	0 0.0%	0 0.0%	100% 0.0%
	Sadness	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	6 14.3%	0 0.0%	100% 0.0%
	Surprise	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	6 14.3%	100% 0.0%
			100% 0.0%	100% 0.0%	100% 0.0%	100% 0.0%	83.3% 16.7%	100% 0.0%	100% 0.0%
		Target Class							

Table 21 The confusion matrix of the ResNet101 model on the test set of FER2013

		Confusion Matrix							
Output Class	Angry	84 5.8%	5 0.3%	37 2.6%	24 1.7%	17 1.2%	27 1.9%	11 0.8%	41.0% 59.0%
	Disgust	4 0.3%	6 0.4%	2 0.1%	2 0.1%	0 0.0%	2 0.1%	0 0.0%	37.5% 62.5%
	Fear	8 0.6%	1 0.1%	48 3.3%	5 0.3%	16 1.1%	26 1.8%	6 0.4%	43.6% 56.4%
	Happy	17 1.2%	5 0.3%	18 1.3%	285 19.8%	27 1.9%	29 2.0%	11 0.8%	72.7% 27.3%
	Neutral	35 2.4%	3 0.2%	36 2.5%	16 1.1%	139 9.7%	42 2.9%	10 0.7%	49.5% 50.5%
	Sad	32 2.2%	1 0.1%	43 3.0%	18 1.3%	35 2.4%	112 7.8%	4 0.3%	45.7% 54.3%
	Surprised	12 0.8%	1 0.1%	21 1.5%	5 0.3%	13 0.9%	11 0.8%	124 8.6%	66.3% 33.7%
			43.8% 56.3%	27.3% 72.7%	23.4% 76.6%	80.3% 19.7%	56.3% 43.7%	45.0% 55.0%	74.7% 25.3%
		Target Class							

Table 22 Misclassified FER2013 images (Original Class–Predicted Class)



CNN model, their fusion into a single vector, and then the SVM-based emotion classification of the vector. Based on the results obtained by the three models on the four datasets, we can notice that they are complementary. Therefore, it could be beneficial to combine the features in order to meet the objective of tackling the shortcomings of one model through the performance of the other models. This fact led us to suggest fusing features extracted from the three models and then fed them to a supervised classifier. This observation was confirmed by the experimental results after extraction and concatenation of the features which gave the best recognition rates for the three models. Consequently, the use of mixed feature from the three models has considerably improved the overall recognition rate. In fact, the experiments performed using the concatenated features enabled us to achieve an overall recognition rate of 97.62% on JAFFE, 98.80% on CK+, 88.20% on SFEW_2.0, and 94.01% on FER2013. To evaluate the overall performance of the proposed method, the confusion matrices on the three datasets are illustrated in the Tables 23, 24, 25 and 26.

The process of fusing the resulting feature vectors of JAFFE led the proposed FER system to recognize the Sadness emotion perfectly by reaching 100% of recognition rate for this emotional class, while it was recognized at 72.23% on average. Similarly, it reached 100% for Anger and Disgust emotions. The fusion has also led to increase the recognition rate of the Neutral emotion to 83.3%, and to have a recognition rate of 100% for the other emotions. On the CK+ dataset, the resulting feature vector improves all the emotions' recognition rates. This fact is reflected by the increase in the overall accuracy rate to 98.8%. The best recognition

Table 24 Confusion matrix of the proposed method on the CK+ dataset

		Confusion Matrix							
Output Class	Angry	170 15.7%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	100% 0.0%
	Disgust	0 0.0%	134 12.4%	0 0.0%	0 0.0%	2 0.2%	0 0.0%	0 0.0%	98.5% 1.5%
	Fear	0 0.0%	0 0.0%	92 8.5%	0 0.0%	1 0.1%	0 0.0%	0 0.0%	98.9% 1.1%
	Happy	0 0.0%	0 0.0%	0 0.0%	211 19.5%	0 0.0%	0 0.0%	0 0.0%	100% 0.0%
	Neutral	1 0.1%	1 0.1%	1 0.1%	1 0.1%	172 15.9%	0 0.0%	5 0.5%	95.0% 5.0%
	Sad	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	89 8.2%	0 0.0%	100% 0.0%
	Surprised	0 0.0%	0 0.0%	0 0.0%	0 0.0%	1 0.1%	0 0.0%	201 18.6%	99.5% 0.5%
			99.4% 0.6%	99.3% 0.7%	98.9% 1.1%	99.5% 0.5%	97.7% 2.3%	100% 0.0%	97.6% 2.4%
		Target Class							

Table 25 Confusion matrix of the proposed method on the SFEW_2.0 dataset

		Confusion Matrix							
Output Class	Angry	44 17.9%	2 0.8%	1 0.4%	0 0.0%	2 0.8%	0 0.0%	2 0.8%	86.3% 13.7%
	Disgust	1 0.4%	10 4.1%	0 0.0%	0 0.0%	0 0.0%	1 0.4%	0 0.0%	83.3% 16.7%
	Fear	0 0.0%	0 0.0%	22 8.9%	0 0.0%	0 0.0%	0 0.0%	1 0.4%	95.7% 4.3%
	Happy	0 0.0%	0 0.0%	1 0.4%	44 17.9%	0 0.0%	2 0.8%	1 0.4%	91.7% 8.3%
	Neutral	1 0.4%	1 0.4%	0 0.0%	1 0.4%	40 16.3%	4 1.6%	0 0.0%	85.1% 14.9%
	Sad	0 0.0%	1 0.4%	0 0.0%	2 0.8%	2 0.8%	34 13.8%	1 0.4%	85.0% 15.0%
	Surprise	1 0.4%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	1 0.4%	23 9.3%	92.0% 8.0%
		93.6% 6.4%	71.4% 28.6%	91.7% 8.3%	93.6% 6.4%	90.9% 9.1%	81.0% 19.0%	82.1% 17.9%	88.2% 11.8%
		Target Class							

rate of 97.8% achieved by GoogleNet model for the Sadness emotion reached 100% using the concatenated vector. The lowest recognition rate of the Neutral class, which had a mean value of 76.7%, reached 97.7%, whereas for Surprise emotion it had an increase up to 97.6%. An average of 99.4% has been achieved for the Anger, Disgust, Fear, and Happiness emotions. Regarding the second type of datasets (in-wild-conditions), as discussed before, for the SFEW_2.0 dataset, which is more challenging than the other facial expressions datasets due to the complexity of background and the natural situation of human faces, we note a striking improvement in

Table 26 Confusion matrix of the proposed method on the FER2013 dataset

		Confusion Matrix							
Output Class	Angry	178 12.4%	5 0.3%	3 0.2%	1 0.1%	2 0.1%	3 0.2%	1 0.1%	92.2% 7.8%
	Disgust	0 0.0%	15 1.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	100% 0.0%
	Fear	5 0.3%	1 0.1%	187 13.0%	1 0.1%	2 0.1%	3 0.2%	9 0.6%	89.9% 10.1%
	Happy	1 0.1%	0 0.0%	1 0.1%	349 24.3%	0 0.0%	0 0.0%	0 0.0%	99.4% 0.6%
	Neutral	0 0.0%	0 0.0%	4 0.3%	3 0.2%	229 15.9%	7 0.5%	0 0.0%	94.2% 5.8%
	Sad	8 0.6%	1 0.1%	9 0.6%	1 0.1%	14 1.0%	236 16.4%	0 0.0%	87.7% 12.3%
	Surprised	0 0.0%	0 0.0%	1 0.1%	0 0.0%	0 0.0%	0 0.0%	156 10.9%	99.4% 0.6%
		92.7% 7.3%	68.2% 31.8%	91.2% 8.8%	98.3% 1.7%	92.7% 7.3%	94.8% 5.2%	94.0% 6.0%	94.0% 6.0%
		Target Class							

the emotion recognition rate for this dataset from an average of 57.46% to 88.2% with a considerable increase of 30.74%. For the second dataset in-the-wild (FER2013), which is known to be one of the most challenging dataset in emotion recognition domain as it contains images of cartoons and emojis in addition to the human facial images, the recognition rate using the concatenated vector has been clearly improved compared to those of each model apart. Indeed, an augmentation of the lowest accuracy of the Disgust emotion from 18.2 to 68.2%, and of the Fear emotion from 23.4 to 91.2% were recorded. All the other recognition rates have been increased to an average of 92%. The Happy emotion was the most recognized emotion with an accuracy of 98.3%. Many other challenging factors in facial emotion recognition can also reduce the emotion recognition rate, in particular when the images are taken in-the-wild conditions. It is worth mentioning that this type of images is different of images taken in laboratory conditions. In-the-wild conditions, there are different head poses because the individuals are in movement, and the distance between persons and the camera is variable. Contrary to the in controlled conditions where the persons are in front of the camera with the same distance, and vertical head pose. These challenging factors were overcome through the concatenation of relevant features of the three models. Assembling the features has reduced the error rate compared to each model separately and has remarkably improved the overall recognition rate especially for the in-the-wild datasets. The limitations of the single models have been relatively covered by the union of the three models into a global one able to predict more precisely the emotions. Each model has correctly classified a set of

Table 27 Comparison between the FER accuracy of the proposed method and the ones recorded by relevant methods from the state-of-the-art on the following datasets: (a) JAFFE, (b) CK+, (c) SFEW_2.0, (d) FER2013

(a)	
Studies	JAFFE
Mohan et al. [42]	97%
Wu et al. [73]	94.01%
Salman et al. [75]	78.57%
Kim et al. [76]	91.27%
Zhang et al. [52]	92.30%
Hung et al. [77]	90.97%
Jain et al. [78]	95.23%
Xie and Hu [79]	94.75%
Gogic et al. [57]	85.88%
Minaee et al. [9]	92.80%
Ravi and Yadhukrishna [72]	77.27%
Siam et al. [66]	88%
Proposed method	97.62%
(b)	
Studies	CK+
Wu et al. [73]	91%
Shao and Qian [80]	95.29%
Salman et al. [75]	96.92%
Kim et al. [76]	96.46%
Zhang et al. [52]	98.50%
Jain et al. [78]	93.24%
Xie and Hu [79]	93.46%
Gogic et al. [57]	96.48%
Umer et al. [81]	97.69%
Minaee et al. [9]	98.00%
Ravi and Yadhukrishna [72]	97.32%
Shaees et al. [51]	98.30% for the hybrid approach and 90.1% for the transfer learning feature-based approach
Siam et al. [66]	94%
Proposed method	98.80%
(c)	
Studies	SFEW_2.0
Wu et al. [73]	49.02%
Gogic et al. [57]	49.31%
Yan et al. [82]	53.10%
Alreshidi and Ullah [58]	57.70%
Saurav et al. [83]	59.16%
Zhou et al. [84]	52.98%
Proposed method	88.20%

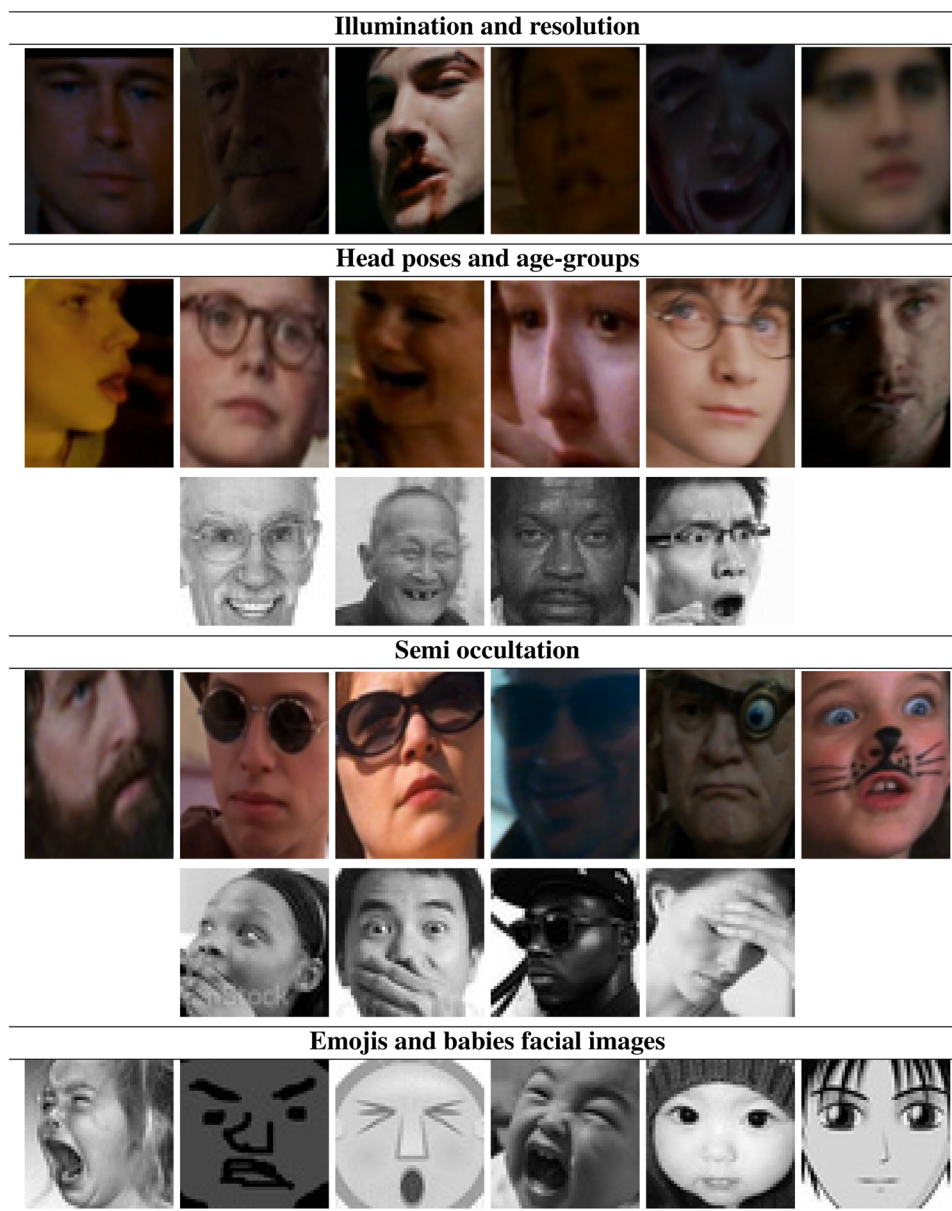
Table 27 continued

(d)	
Studies	FER2013
Minaee et al. [9]	70.02%
Saurav et al. [67]	72.77%
Devi Bodapati et al. [74]	69.57%
Liang et al. [85]	72.81%
Mohan et al. [42]	79%
Proposed method	94.02%

images that is different from the set correctly classified by the other model. The idea of feature concatenation applied in this work was able to ensure the maximum number of images correctly classified by the three models, especially in the case of in-the-wild datasets such as FER2013 and SFEW_2.0. This justifies the performance increase from 50% to 80%. Overall, the obtained results show that using the complementarity of several deep learning models and extracting features from different models can counteract the difficulties of capturing facial emotions in-the-wild.

4.7 Comparison of the suggested method with relevant methods from the state-of-the-art

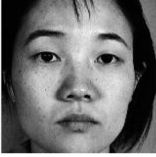









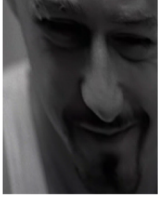

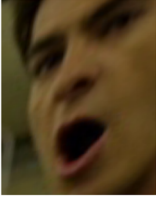
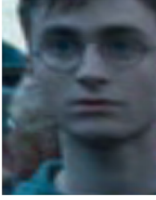
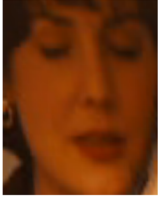





This section presents a comparison of the proposed FER method with relevant emotion recognition methods from the literature. For fair comparison, we ascertained that the compared methods are using deep learning, transfer learning and CNN architectures, in addition to be validated on the same datasets as those of this study. The comparison results on the JAFFE, the CK+, the SFEW_2.0, and the FER2013 datasets are summarized in Table 27 (a, b, c, and d, respectively). It is clear that the proposed method has outperformed all the compared methods on all datasets. For instance, the proposed method is outperforming the method presented in [72] on the JAFFE dataset with a difference of 20.35%. Furthermore, for the case of the CK+ dataset, the proposed method accuracy is better than the accuracies reached by the method in [73] with a difference of 7.8%, and the one in [51] with a difference of 8.7%. The average increase of accuracy compared to the other methods varies from 0.3 to 5.56%. Likewise, while investigating the SFEW_2.0 dataset, the performances of the proposed method have outperformed those obtained by relevant state-of-the-art methods with a considerable difference of 39.18%, which represents the highest increase among all the studied datasets. Similarly, the validation of proposed method on the FER2013 dataset allows an improvement of results comparing to recent studies. In fact, the improvement of the recognition rate reached 24.45% compared to [74], and 15.02% compared to the best accuracy obtained by [42].

Table 28 Sample of challenging in-the-wild conditions

Overall, the proposed method has reached better rates with regards to all the compared methods for the four used datasets. In particular, there is a substantial increase in the recognition of spontaneous emotions within the SFEW_2.0 and the FER2013 datasets. This was expected since the proposed method is designed to take advantage of the complementarity of deep learning models, especially for the in-the-wild context. We can also notice that the proposed model misclassifications concern above all the classes with less samples compared with other classes, for instance, the “Disgust” class of the SFEW_2.0 and the FER2013 datasets, which include only 14 and 22 samples, respectively. This emotion has been recognized at 71.4% for the SFEW_2.0

dataset, and at 68.2% for the FER2013 dataset. Thus, seven images have not been correctly classified from the FER2013 dataset, and just four facial images have not been correctly classified from the SFEW_2.0 dataset, particularly for the images where the expressions are not accentuated such as the ones presented in Table 29(c') and the two last images in Table 29(c''). Likewise for the FER2013 dataset, in Table 29(d) the images do not strongly express the emotions. Under those circumstances, it is difficult even for the human being to capture the specific emotion. Table 29 gathers qualitative results for some examples from the four datasets. For the JAFFE dataset, the only misclassification made by the model was for a facial image where the “Fear” expression is

Table 29 A sample of misclassified images by the proposed method: Original Class (green)–Predicted Class (red)

		JAFFE				
(a)		Neutral				
						
		Fear				
		CK+				
(b)		Disgust	Happy	Surprise	Surprise	
						
		Neutral	Neutral	Neutral	Neutral	
		Neutral		Neutral		
						
		Disgust		Fear		
		SFEW_2.0				
(c')		Disgust	Disgust	Disgust	Disgust	
						
		Anger	Neutral	Sad	Anger	
		Anger	Neutral	Sad	Happy	
(c'')						
		Surprise	Sad	Neutral	Sad	
			FER2013			
	(d)		Disgust	Disgust	Sad	Fear
						
		Sad	Angry	Neutral	Neutral	

too similar to the “Neutral” one. Considering all the expressions of the corresponding subject, we notice that they are too similar and are not too expressive, which explains the error in Table 29(a). For the case of the CK+ dataset, misclassifications also occur as shown in Table 29(b). Indeed, because of the similarities between facial expressions, we have encountered errors where some images from different classes have been classified as “Neutral” emotion owing to similarities between facial expressions. Furthermore, the misclassified images are the first instances of the image sequence that describes the emotion, where the expression of the emotion has not yet appeared. These images are equivalent to the first three images of each sequence which constituted the Neutral class. Concerning the second type of datasets that include images captured in the real world, some facial emotions have been incorrectly predicted by the model for the SFEW_2.0 dataset, as shown in Table 29(c). We may attribute the failure of emotion recognition into cases where there are different emotions intensities, various poses and lighting conditions of movie scenes, and other uncontrolled conditions that exist in this dataset. Mainly, we mention resolution variations, different age groups, occlusion, in addition to the previous ones. The dataset contains even images with more than two challenges at a time. The different challenges are illustrated in Table 28. These images are delicate even for the human being to classify. Nevertheless, the proposed scheme was able to achieve 88.20% accuracy thanks to the quality and the complementarity of the relevant features extracted from each model. As well as we know, this result has never been achieved before, and this is outstanding for the challenging issue of the in-the-wild FER. Passing to the FER2013 dataset, some images have not been correctly classified as presented in Table 29(d). The degradation of results in these cases is due to the unbalanced distribution of the number of samples per class. For instance, we find the “Disgust” class containing 111 samples, while the “Happy” class includes 1774 samples. In addition to the uncontrolled conditions existing within the other datasets, FER2013 is characterized by other specific uncontrolled constraints, such as the high number of babies and children’s facial images, the different skin colors and features, and the presence of cartoon images and emojis (Table 28). To sum up, most images that were misclassified by the individual models have been correctly classified and recognized using the deep features extracted and concatenated into a single feature vector. Moreover, images that have been misclassified after the feature concatenation were misclassified by at least two of the three models and in most cases by all three models. These images are either very challenging (incorporating occlusions and extreme head pose deviation) or images with very low level intensity of the expressions. The concatenation of deep features while choosing the suitable layers was generally able to raise the problem of false classification by individual models. In fact, the concatenation

has filled the lack of one model by the other models, through the dynamic selection of the more relevant layers that contain the most discriminant features.

Overall, according to the experimental results, using a multichannel CNN method based on deep learning techniques on the well-known CK+, JAFFE, FER2013, and SFEW_2.0 datasets, the proposed method shows high recognition accuracy thanks to the richness of each selected pre-trained model in this study (VGG19, GoogleNet, ResNet101) as well as to the relevance of the deep features extracted from each one. Besides, the freezing of the layers applied on a personalized level relative to the depth of the pre-trained model led us to gain time and quality of extracted features. Indeed, the recorded execution time by the proposed method for the test phase is 16.537 ms, 47.200 ms, 36.463 ms, and 36.149 ms for the JAFFE, the CK+, the FER2013, and the SFEW_2.0 datasets, respectively, using the following hardware configuration: Intel(R) i7 9th generation CPU, with NVIDIA GeForce RTX2060 GPU, and 16GB RAM. This computational cost is among the best costs recorded by the state-of-the-art works that have been tested on the same datasets. For instance, although that more powerful hardware configuration (16GB GPU RAM, 2560 Cuda cores, 256-bit memory interface, GDDR5X as memory type NVIDIA Quadro P5000) has been used in [42], execution times of 402.6 ms, 569.7 ms and 1161.2 ms have been recorded for the JAFFE, the CK+ and the FER2013 datasets, respectively. However, in [75] (*resp.* [66]), an average execution time of 60 ms (*resp.* 34 ms) have been reported when testing on the three datasets using almost a similar hardware configuration then the one used in our work. Thus, the proposed method has proved to be cost-efficient in terms of computational time with an average of 34 ms, what represents half of the required time for the work of [75] while being competitive compared to [66].

5 Conclusion

In order to depict facial emotions more accurately, we have proposed a robust and computationally efficient method based on multichannel deep features extraction and concatenation. The proposed FER method is based upon two dual deep learning networks: the first one is dedicated to deep features’ extraction from three CNN models, while the second one is used for feature selection and concatenation. The validity of the proposed method has been assessed on four widely used FER datasets. The investigated datasets present various types of emotions: posed and spontaneous emotions, in the laboratory-conditions and in the wild-conditions. The first-line experiments performed in this study, using the three pre-trained CNN models, led to two findings. Firstly, the three used CNN models are highly effective to capture human

facial emotions, and secondly, these models do not have the same weaknesses and strengths regarding the recognition of the different classes of emotions. Therefore, the main challenge of the suggested method is to consider deep features coming from multiple multichannel convolutional neural networks, thereby leading to getting the most out of the models' performances by the complementarity. The main objective is to achieve better results than those of each model applied separately, notably for the case of in-the-wild environments. The experimental study of the proposed method for emotion recognition has been divided into two parts. The first one has presented the results of the DL as extractor. It highlights the efficiency of the geometric DA techniques, which allowed increasing the amount of training images. This type of techniques enabled to improve the training relevance of deep data remaining after applying freezing weights. These results also emphasize the efficiency of features extracted from these remaining layers, and the quality of data existing in each model. The second part has shown the results of the DL as transformer. It displays the final recognition rates of the output of the multichannel convolutional neural network. This DL aims to possess a single emotion prediction vector for each dataset. Thus, the final vector encompasses the most relevant features selected from rich layers what has helped to improve the final accuracy. In summary, the suggested method outperforms many relevant state-of-the-art methods, in addition to all the single model-based methods. It achieved 97.62% for the JAFFE dataset, and 98.80% for the CK+ dataset, while obtaining 88.20% for the SFEW_2.0 dataset, and 94.01% for the FER2013 dataset. According to the experimental results and to the comparative analysis with reference to several state-of-the-art works, we point out that the obtained results in this work outperform those in the literature for four datasets, especially for the FER2013 and the SFEW_2.0 datasets. Those latter elucidated a significant higher recognition rate of 20%, and 34%, respectively, on average compared to previous recognition rates. This confirms the ability of the concatenated vector, formed by heterogeneous deep features extracted from the three CNN models, to enhance the accuracy of emotion recognition, particularly for the in-the-wild datasets where the enhancement has been remarkable. Furthermore, the proposed method can be ameliorated in future by investigating the use of action units and face landmarks in order to improve the recognition rate of image faces with non-accentuated expressions.

Declarations

Conflict of interest The authors have no relevant financial or non-financial interests to disclose.

References

1. Plutchik, R.: *Emotion, a Psychoevolutionary Synthesis*. Harper & Row, New York (1980)
2. Bericat, E.: The sociology of emotions: four decades of progress. *Curr. Sociol.* **64**(3), 491–513 (2016)
3. Shouse, E.: Feeling, emotion, affect. *M/C Journal* **8**(6) (2005)
4. Darwin, C., Prodger, P.: *The Expression of the Emotions in Man and Animals*. Oxford University Press, Oxford (1998)
5. Damasio, A., Blanc, M.: *L'erreur de Descartes: la raison des émotions*. Odile Jacob, France (2006)
6. Karnati, M., Seal, A., Yazidi, A., Krejcar, O.: Lienet: a deep convolution neural networks framework for detecting deception. *IEEE Trans. Cognit. Dev. Syst.* **14**(3), 971–984 (2022)
7. Hua, W., Dai, F., Huang, L., Xiong, J., Gui, G.: Hero: human emotions recognition for realizing intelligent internet of things. *IEEE Access* **7**, 24321–24332 (2019)
8. Khorsheed, J.A., Yurtkan, K.: Analysis of local binary patterns for face recognition under varying facial expressions. In: *Signal Processing and Communication Application Conference (SIU)*, pp. 2085–2088 (2016)
9. Minaee, S., Minaei, M., Abdolrashidi, A.: Deep-emotion: facial expression recognition using attentional convolutional network. *Sensors* **21**(9), 3046 (2021)
10. Lyons, M., Kamachi, M., Gyoba, J.: The Japanese female facial expression (JAFFE) dataset (1998)
11. Kanade, T., Cohn, J.F., Tian, Y.: Comprehensive database for facial expression analysis. In: *IEEE International Conference on Automatic Face and Gesture Recognition (Cat. No. PR00580)*, pp. 46–53 (2000)
12. Lucey, P., Cohn, J.F., Kanade, T., Saragih, J., Ambadar, Z., Matthews, I.: The extended Cohn-Kanade dataset (CK+): a complete dataset for action unit and emotion-specified expression. In: *IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*, pp. 94–101 (2010)
13. Zhao, G., Huang, X., Taini, M., Li, S.Z., Pietikäinen, M.: Facial expression recognition from near-infrared videos. *Image Vis. Comput.* **29**(9), 607–619 (2011)
14. Mollahosseini, A., Hasani, B., Mahoor, M.H.: Affectnet: a database for facial expression, valence, and arousal computing in the wild. *IEEE Trans. Affect. Comput.* **10**(1), 18–31 (2017)
15. Dhall, A., Goecke, R., Lucey, S., Gedeon, T.: Collecting large, richly annotated facial-expression databases from movies. *IEEE Multimed.* **19**(3), 34–41 (2012)
16. Dhall, A., Goecke, R., Joshi, J., Sikka, K., Gedeon, T.: Emotion recognition in the wild challenge 2014: baseline, data and protocol. In: *International Conference on Multimodal Interaction*, pp. 461–466 (2014)
17. Goodfellow, I.J., Erhan, D., Carrier, P.L., Courville, A., Mirza, M., Hamner, B., Cukierski, W., Tang, Y., Thaler, D., Lee, D.-H. et al.: Challenges in representation learning: A report on three machine learning contests. In: *International Conference on Neural Information Processing (ICONIP)*, pp. 117–124 (2013)
18. Zhang, L., Verma, B., Tjondronegoro, D., Chandran, V.: Facial expression analysis under partial occlusion: a survey. *ACM Comput. Surv.* **51**(2), 25 (2018)
19. Tian, Y.-L., Kanade, T., Cohn, J.F.: *Facial expression analysis*. In: *Handbook of Face Recognition*, pp. 247–275. Springer (2005)
20. Mollahosseini, A., Hasani, B., Mahoor, M.H.: Affectnet: a database for facial expression, valence, and arousal computing in the wild. *IEEE Trans. Affect. Comput.* **10**(1), 18–31 (2019)
21. Dou, S., Feng, Z., Yang, X., Tian, J.: Real-time multimodal emotion recognition system based on elderly accompanying robot. *J. Phys.: Conf. Ser.* **1453**(1), 012093 (2020)

22. Khorrami, P., Paine, T., Huang, T.: Do deep neural networks learn facial action units when doing expression recognition? In: IEEE International Conference on Computer Vision Workshop (ICCVW), pp. 19–27 (2015)
23. Ben Fredj, H., Bouguezzi, S., Souani, C.: Face recognition in unconstrained environment with CNN. *Vis. Comput.* **37**(2), 217–226 (2021)
24. Kumar, A., Kaur, A., Kumar, M.: Face detection techniques: a review. *Artif. Intell. Rev.* **52**, 927–948 (2019)
25. Zhang, K., Zhang, Z., Li, Z., Qiao, Y.: Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Process. Lett.* **23**(10), 1499–1503 (2016)
26. King, D.E.: Dlib-ml: a machine learning toolkit. *J. Mach. Learn. Res.* **10**, 1755–1758 (2009)
27. Wirdiani, N.A., Lattifia, T., Supadma, I.K., Mahar, B.K., Taradhita, D.N., Fahmi, A.: Real-time face recognition with eigenface method. *Int. J. Image Graph. Signal Process.* **11**(11), 1–9 (2019)
28. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR) (2001)
29. Li, S., Deng, W.: Deep facial expression recognition: a survey. *IEEE Trans. Affect. Comput.* **13**(3), 1195–1215 (2022)
30. Lopes, A.T., De Aguiar, E., Oliveira-Santos, T.: A facial expression recognition system using convolutional networks. In: SIBGRAPI Conference on Graphics, Patterns and Images, pp. 273–280 (2015)
31. Li, K., Jin, Y., Akram, M.W., Han, R., Chen, J.: Facial expression recognition with convolutional neural networks via a new face cropping and rotation strategy. *Vis. Comput.* **36**(2), 391–404 (2020)
32. Ramos, A.L.A., Dadiz, B.G., Santos, A.B.G.: Classifying emotion based on facial expression analysis using Gabor filter: a basis for adaptive effective teaching strategy. In: Computational Science and Technology, vol. 603, pp. 469–479. Springer (2020)
33. Slimani, K., Kas, M., El Merabet, Y., Messoussi, R., Ruichek, Y.: Facial emotion recognition: A comparative analysis using 22 LBP variants. In: Mediterranean Conference on Pattern Recognition and Artificial Intelligence, pp. 88–94 (2018)
34. Kumar, P., Happy, S., Routray, A.: A real-time robust facial expression recognition system using hog features. In: International Conference on Computing, Analytics and Security Trends (CAST), pp. 289–293 (2016)
35. Liu, X., Cheng, X., Lee, K.: GA-SVM-based facial emotion recognition using facial geometric features. *IEEE Sens. J.* **21**(10), 11532–11542 (2020)
36. Zhang, H., Su, W., Wang, Z.: Weakly supervised local-global attention network for facial expression recognition. *IEEE Access* **8**, 37976–37987 (2020)
37. Zhang, S., Zhao, X., Lei, B.: Facial expression recognition based on local binary patterns and local fisher discriminant analysis. *WSEAS Trans. Signal Process.* **8**(1), 21–31 (2012)
38. Abdulrahman, M., Eleyan, A.: Facial expression recognition using support vector machines. In: Signal Processing and Communications Applications Conference (SIU), pp. 276–279 (2015)
39. Alshamsi, H., Kepuska, V., Meng, H.: Automated facial expression recognition app development on smart phones using cloud computing. In: Annual Ubiquitous Computing, Electronics and Mobile Communication Conference (UEMCON), pp. 577–583 (2017)
40. Gite, B., Nikhal, K., Palnak, F.: Evaluating facial expressions in real time. In: Intelligent Systems Conference (IntelliSys), pp. 849–855 (2017)
41. Agrawal, A., Mittal, N.: Using CNN for facial expression recognition: a study of the effects of kernel size and number of filters on accuracy. *Vis. Comput.* **36**(2), 405–412 (2020)
42. Mohan, K., Seal, A., Krejcar, O., Yazidi, A.: Fer-net: Facial expression recognition using deep neural net. *Neural Comput. Appl.* **33**(15), 9125–9136 (2021)
43. Siqueira, H., Magg, S., Wermter, S.: Efficient facial feature learning with wide ensemble-based convolutional neural networks. In: AAAI Conference on Artificial Intelligence, pp. 5800–5809 (2020)
44. Fasel, B.: Robust face analysis using convolutional neural networks. In: International Conference on Pattern Recognition (ICPR), pp. 40–43 (2002)
45. Mohan, K., Seal, A., Krejcar, O., Yazidi, A.: Facial expression recognition using local gravitational force descriptor-based deep convolution neural networks. *IEEE Trans. Instrum. Measur.* **70**, 1–12 (2021)
46. Liu, P., Han, S., Meng, Z., Tong, Y.: Facial expression recognition via a boosted deep belief network. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1805–1812 (2014)
47. Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., Manzagol, P.-A., Bottou, L.: Stacked denoising autoencoders: learning useful representations in a deep network with a local denoising criterion. *J. Mach. Learn. Res.* **11**(12), 3371–3408 (2010)
48. Fan, Y., Lu, X., Li, D., Liu, Y.: Video-based emotion recognition using CNN-RNN and C3D hybrid networks. In: ACM International Conference on Multimodal Interaction, pp. 445–450 (2016)
49. Lai, Y.-H., Lai, S.-H.: Emotion-preserving representation learning via generative adversarial network for multi-view facial expression recognition. In: IEEE International Conference on Automatic Face & Gesture Recognition, pp. 263–270 (2018)
50. Hazarika, D., Poria, S., Zimmermann, R., Mihalcea, R.: Conversational transfer learning for emotion recognition. *Inf. Fusion* **65**, 1–12 (2021)
51. Shaees, S., Naeem, H., Arslan, M., Naeem, M.R., Ali, S.H., Aldabbas, H.: Facial emotion recognition using transfer learning. In: International Conference on Computing and Information Technology (ICCIT), pp. 1–5 (2020)
52. Zhang, H., Huang, B., Tian, G.: Facial expression recognition based on deep convolution long short-term memory networks of double-channel weighted mixture. *Pattern Recognit. Lett.* **131**, 128–134 (2020)
53. Chen, L., Wu, M., Pedrycz, W., Hirota, K.: Deep sparse autoencoder network for facial emotion recognition. In: Emotion Recognition and Understanding for Emotional Human–Robot Interaction Systems, Studies in Computational Intelligence, pp. 25–39. Springer International Publishing (2021)
54. Liu, Y., Zhang, X., Lin, Y., Wang, H.: Facial expression recognition via deep action units graph network based on psychological mechanism. *IEEE Trans. Cognit. Dev. Syst.* **12**(2), 311–322 (2019)
55. Fan, X., Tjahjadi, T.: Fusing dynamic deep learned features and handcrafted features for facial expression recognition. *J. Vis. Commun. Image Represent.* **65**, 102659 (2019)
56. Sun, X., Lv, M.: Facial expression recognition based on a hybrid model combining deep and shallow features. *Cognit. Comput.* **11**(4), 587–597 (2019)
57. Gogic, I., Manhart, M., Pandžic, I.S., Ahlberg, J.: Fast facial expression recognition using local binary features and shallow neural networks. *Vis. Comput.* **36**, 97–112 (2020)
58. Alreshidi, A., Ullah, M.: Facial emotion recognition using hybrid features. *Informatics* **7**(1), 6 (2020)
59. Dhall, A., Ramana Murthy, O., Goecke, R., Joshi, J., Gedeon, T.: Video and image based emotion recognition challenges in the wild: EmotiW 2015. In: ACM International Conference on Multimodal Interaction, pp. 423–426 (2015)
60. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: International Conference on Learning Representations (ICLR) (2015)
61. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1–9 (2015)

62. Li, Y., Lyu, S.: Exposing deepfake videos by detecting face warping artifacts. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 46–52 (2018)
63. Boughanem, H., Ghazouani, H., Barhoumi, W.: Towards a deep neural method based on freezing layers for in-the-wild facial emotion recognition. In: *IEEE/ACS International Conference on Computer Systems and Applications (AICCSA)*, pp. 1–8 (2021)
64. Ahmed, T.U., Hossain, S., Hossain, M.S., ul Islam, R., Andersson, K.: Facial expression recognition using convolutional neural network with data augmentation. In: *Joint International Conference on Informatics, Electronics & Vision (ICIEV) and International Conference on Imaging, Vision & Pattern Recognition (icIVPR)*, pp. 336–341 (2019)
65. Simard, P.Y., Steinkraus, D., Platt, J.C.: Best practices for convolutional neural networks applied to visual document analysis. In: *International Conference on Document Analysis and Recognition (ICDAR)*, pp. 958–963 (2003)
66. Siam, A.I., Soliman, N.F., Algarni, A.D., El-Samie, A., Fathi, E., Sedik, A.: Deploying machine learning techniques for human emotion detection. *Comput. Intell. Neurosci.* **2022**, 8032673 (2022)
67. Saurav, S., Gidde, P., Saini, R., Singh, S.: Dual integrated convolutional neural network for real-time facial expression recognition in the wild. *Vis. Comput.* **38**(3), 1083–1096 (2022)
68. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. *Commun. ACM* **60**(6), 84–90 (2017)
69. Vo, A., Nguyen, B.T.: Facial expression recognition based on salient regions. In: *International Conference on Green Technology and Sustainable Development (GTSD)*, pp. 739–743 (2018)
70. Huang, Y., Chen, F., Lv, S., Wang, X.: Facial expression recognition: a survey. *Symmetry* **11**(10), 1189 (2019)
71. Dhall, A., Goecke, R., Lucey, S., Gedeon, T.: Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark. In: *IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, pp. 2106–2112 (2011)
72. Ravi, R., Yadhukrishna, S., Prithviraj, R.: A face expression recognition using CNN & LBP. In: *International Conference on Computing Methodologies and Communication (ICCMC)*, pp. 684–689 (2020)
73. Wu, M., Su, W., Chen, L., Liu, Z., Cao, W., Hirota, K.: Weight-adapted convolution neural network for facial expression recognition in human–robot interaction. *IEEE Transactions on Systems, Man, and Cybernetics: Systems* **51**(3), 1473–1484 (2021)
74. Bodapati, J.D., Srilakshmi, U., Veeranjanyulu, N.: Fernet: a deep CNN architecture for facial expression recognition in the wild. *J. Inst. Eng. (India): Ser. B* **103**(2), 439–448 (2022)
75. Salmam, F.Z., Madani, A., Kissi, M.: Fusing multi-stream deep neural networks for facial expression recognition. *Signal Image Video Process.* **13**(3), 609–616 (2019)
76. Kim, J.-H., Kim, B.-G., Roy, P.P., Jeong, D.-M.: Efficient facial expression recognition algorithm based on hierarchical deep neural network structure. *IEEE Access* **7**, 41273–41285 (2019)
77. Hung, J.C., Lin, K.-C., Lai, N.-X.: Recognizing learning emotion based on convolutional neural networks and transfer learning. *Appl. Soft Comput.* **84**, 105724 (2019)
78. Jain, D.K., Shamsolmoali, P., Sehdev, P.: Extended deep neural network for facial emotion recognition. *Pattern Recognit. Lett.* **120**, 69–74 (2019)
79. Xie, S., Hu, H.: Facial expression recognition using hierarchical features with deep comprehensive multipatches aggregation convolutional neural networks. *IEEE Trans. Multimed.* **21**(1), 211–220 (2018)
80. Shao, J., Qian, Y.: Three convolutional neural network models for facial expression recognition in the wild. *Neurocomputing* **355**, 82–92 (2019)
81. Umer, S., Rout, R.K., Pero, C., Nappi, M.: Facial expression recognition with trade-offs between data augmentation and deep learning features. *J. Ambient. Intell. Humaniz. Comput.* **13**(2), 721–735 (2022)
82. Yan, K., Zheng, W., Zhang, T., Zong, Y., Tang, C., Lu, C., Cui, Z.: Cross-domain facial expression recognition based on transductive deep transfer learning. *IEEE Access* **7**, 108906–108915 (2019)
83. Saurav, S., Saini, R., Singh, S.: Emnet: a deep integrated convolutional neural network for facial emotion recognition in the wild. *Appl. Intell.* **51**, 5543–5570 (2021)
84. Zhou, L., Fan, X., Tjahjadi, T., Das Choudhury, S.: Discriminative attention-augmented feature learning for facial expression recognition in the wild. *Neural Comput. Appl.* **34**(2), 925–936 (2022)
85. Liang, X., Xu, L., Zhang, W., Zhang, Y., Liu, J., Liu, Z.: A convolution-transformer dual branch network for head-pose and occlusion facial expression recognition. *Vis. Comput.* (2022)

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

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.



Hadjer Boughanem Algerian Ph.D. student at the University of Tunis El Manar. Her doctoral thesis focuses on emotion's recognition and artificial intelligence. She graduated with a master's degree in image processing and artificial vision from the University of Annaba - Algeria. In addition to a master's degree in multimedia design and integration from the University of Lyon 2 France.



Haythem Ghazouani received B.E. and M.E. degrees in computer science from National School of Computer Science (ENSI), Tunisia, in 2006 and 2007, respectively. He received the Ph.D. in computer science, jointly, from ENSI and University of Montpellier 2 (UM2), France, in 2012. Since 2011, he has been with the National School of Engineering of Carthage, where he is currently assistant professor and head of the computer science department. His main areas of research interest,

within the research team on Intelligent Systems in Imaging and Artificial Vision of the laboratory LIMTIC Laboratory, are image processing and artificial intelligence.



Walid Barhoumi holds a Ph.D. from the National School of Computer Science (Tunisia) and Habilitation from the University of Carthage. He is Associate Professor at the National Engineering School of Carthage (ENICarthage) and a senior researcher at the research team on Intelligent Systems in Imaging and Artificial Vision (SIIVA) of the laboratory LIMTIC, Tunisia.