IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received July 23, 2018, accepted September 7, 2018, date of publication September 26, 2018, date of current version October 17, 2018.

*Digital Object Identifier 10.1109/ACCESS.2018.2870063*

# Physiological Inspired Deep Neural Networks for Emotion Recognition

**PEDRO M. FERREIRA[1,2], FILIPE MARQUES[2], JAIME S. CARDOSO[1,2], (Senior Member, IEEE), AND ANA REBELO[1,3]**

[1]Centre for Telecommunications and Multimedia, INESC TEC, 4200-465 Porto, Portugal
[2]Faculdade de Engenharia da Universidade do Porto, 4200-465 Porto, Portugal
[3]Universidade Portucalense Oporto Portugal, 4200-072 Porto, Portugal

Corresponding author: Pedro M. Ferreira (pmmf@inesctec.pt)

This work was supported in part by the European Regional Development Fund (ERDF) through the Operational Programme for Competitiveness and Internationalisation—COMPETE 2020 Programme, under Project POCI-01-0145-FEDER-006961, and in part by the National Funds through the Portuguese funding agency, Fundação para a Ciência e a Tecnologia (FCT ) under Project UID/EEA/50014/2013, and in part by Ph.D. and BPD under Grant SFRH/BD/102177/2014 and Grant SFRH/BPD/101439/2014.

**ABSTRACT** Facial expression recognition (FER) is currently one of the most active research topics due to its wide range of applications in the human–computer interaction field. An important part of the recent success of automatic FER was achieved thanks to the emergence of deep learning approaches. However, training deep networks for FER is still a very challenging task, since most of the available FER data sets are relatively small. Although *transfer learning* can partially alleviate the issue, the performance of deep models is still below of its full potential as deep features may contain redundant information from the pre-trained domain. Instead, we propose a novel end-to-end neural network architecture along with a well-designed loss function based on the strong prior knowledge that facial expressions are the result of the motions of some facial muscles and components. The loss function is defined to regularize the entire learning process so that the proposed neural network is able to explicitly learn expression-specific features. Experimental results demonstrate the effectiveness of the proposed model in both lab-controlled and wild environments. In particular, the proposed neural network provides quite promising results, outperforming in most cases the current state-of-the-art methods.

**INDEX TERMS** Facial expressions recognition, convolutional neural networks, regularization, domain-knowledge.

## I. INTRODUCTION

Facial expressions (FEs) can be defined as the facial changes in response to a person's internal emotional state, intentions, or social communication [1]. Together with voice, language, hand gestures and body posture, they form a fundamental communication system between humans in social contexts. FEs were introduced as a research field by Charles Darwin in his book "*The Expression of the Emotions in Man and Animals*" [2]. Since then, FEs were established as one of the most important features of human emotion recognition.

In the last few years, automatic facial expression recognition (FER) has attracted much attention due to its wide range of applications, such as human-robot interaction, data-driven animation, interactive games, crowd analytics, biometrics, clinical monitoring and many other human-computer interaction (HCI) systems [1], [3].

Expression recognition is a task that human beings perform daily and effortlessly, but it is not yet easily performed by computers. Although recent methods have demonstrated remarkable performances in highly controlled environments (i.e., high-resolution frontal faces with uniform backgrounds), the automatic FER in real-world scenarios is still a very challenging task [3]. Those challenges are mainly related to different acquisition conditions and to the inter-individual's facial expressiveness variability (see Figure 1a and Figure 1b, respectively). Figure 1b shows six subjects with the angry expression. As illustrated in the figure, the images vary a lot from each other not only in the way that the subjects show their expression, but also in lighting, brightness, viewing angle, pose, position, occlusions and background.

The majority of existing FER systems focus on classifying 6 basic (prototypical) expressions, which have been found to

**FIGURE 1.** Illustration of the main challenges of FER. Those challenges are mainly related to: (a) several physical factors such as pose, viewing angle, occlusions and illumination; and (b) psychological factors such as the inter-individuals facial expressivenesses variability.

**FIGURE 2.** The six basic facial expressions. From left to right: surprise, sadness, fear, anger, disgust and happy.

be universal across cultures and subgroups, namely: happy, surprise, fear, anger, sadness, and disgust (see Figure 2); some systems also recognize the neutral and the contempt expressions [3]. Fewer works follow the dimensional approach, in which the FER is treated as regression problem in a continuous two-dimensional space, usually arousal and valence [4], [5]. It is the example of the research work proposed by Kosti *et al*. [5]. The authors proposed a very complete database that comprises annotations regarding the discrete emotional categories as well as the continuous emotional dimensions. The higher dimensionality of the arousal/valence space potentially allows describing more complex and subtle emotions. However, this richer representation of the expressions is more difficult to use in practice, since the linkage of such dimensional representation to a specific emotion is not straightforward [3]. Other works also attempt to recognize micro-expressions [6]. Micro-expressions are brief involuntary facial expressions that reveal the emotions that a person tries to conceal, especially in high-stakes situations [6]. A very comprehensive and recent survey on FER can be found in [3].

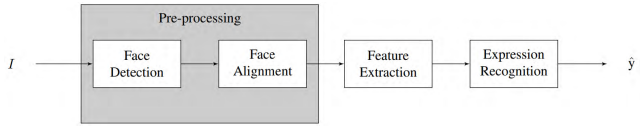



**FIGURE 3.** Diagram of blocks of a typical FER system, where $I$ denotes the input image and $\hat{y}$ represents the predicted FE.

An automatic facial analysis system is typically composed by three main steps: (i) face detection and/or alignment, (ii) feature extraction, and (iii) expression recognition (see Figure 3). Face detection and face alignment are important pre-processing steps for background removal and, then, to rotate or frontalize the face. Effective expression analysis is tightly coupled with the feature extraction step. According to the adopted feature representation, previous FER approaches can be roughly categorized into two main groups: geometric-based methods [7]–[11] and appearance-based methods [12]–[18], [18]–[20]. Geometric-based methods involve, in a first stage, the location of facial landmarks and/or some facial components (e.g., mouth, eyes, nose and eyebrows) and, then, the extraction of geometric features from these fiducial points. Geometric features attempt to measure distances, deformations, curvatures and other geometric properties to represent the face geometry. Appearance-based methods rely on the principle that facial expressions involve change in local texture. Typically, a bank of filters, such as *Local Binary Patterns* [12], *Gabor filters* [13], [14], *Local Gabor Binary Patterns* [15], [16], *Local Phase Quantization* [17], [18], *Scale Invariant Feature Transform* (SIFT) [19], [20], and *Pyramids of Histograms of Gradients* [18], are applied to either the whole face or specific face regions to encode the texture. However, the performance of these hand-crafted feature extraction methods decreases in illumination changes, noise variability, changes in pose, and expression conditions [21]. Another commonly used local feature extraction method for FER is the Local Fisher Discriminant Analysis (LFDA) [22]. In a related work, Kosti *et al*. [5] recently employed the Stepwise Linear Discriminant Analysis (SWLDA) for a robust FER. However, LFDA and SWLDA fail to determine the essential assorted structure when face image space is highly nonlinear.

The recent success of deep learning approaches, particularly those using Convolutional Neural Networks (CNNs), in tasks like object detection and recognition, has been extended to the FER problem. The underlying motivation is to avoid the extraction of hand-crafted features, either geometric- or appearance-based, and the inherent difficulty of designing reliable features to the large inter-individual's facial expressiveness variability. Unlike hand-crafted feature extraction approaches, CNNs are able to automatically learn multiple levels of representations from the data, with higher levels representing more abstract concepts. In general, deep learning approaches became feasible due to two main reasons: (i) the larger amount of data that is currently available in most of the applications, and (ii) the recent advances in GPU technology. The former is crucial for training neural networks with deep architectures without overfitting, whereas the latter is crucial for performing the numerical computations required for the training procedure. However, this is not the case of the FER field, where the availability of large datasets is scarce.

To work around the problem of training high-capacity classifiers on small datasets, previous FER works have mainly resorted to (i) *transfer learning*, where a CNN is typically pre-trained in some domain-related dataset before being fine-tuned to the target dataset; and (ii) *classifier ensembles*, in which an ensemble of CNNs is created in order to combine

their decisions and, hence, reduce the model's variance. However, the gains of *transfer learning* hugely depend on the source-target domain similarity and the availability of an auxiliary large dataset. In addition, the success of *classifier ensembles* requires a wide range of diverse single CNN models.

Inspired by the strong support from physiology and psychology that FEs are the result of the motions of facial muscles [3], [23], [24], a novel end-to-end deep neural network along with a well-designed loss function for FER are proposed. The loss function is defined in a such manner to regularize the entire learning process, so that the proposed model is able to automatically learn expression-specific features. The neural network is composed by three well-designed modules or components, i.e. the *facial-parts component*, the *representation component* and the *classification component*. The purpose of the *facial-parts component* is to regress a relevance map, representing the most important facial regions for the recognition. The relevance map is then used in the *representation component* in order to increase the discriminative ability of the learned features. The result is a model able to explicitly encode expression-specific features by capturing local appearance variations caused by the motion of facial muscles (e.g., frown, grin and glare) and facial components (e.g., eyes, nose, mouth and eyebrows). In addition, according to the level of the available data annotations, different regularization schemes, the so-called fully supervised and weakly supervised regularization schemes, are proposed. The fully supervised regularization scheme is suitable for datasets in which both facial landmarks and expressions are annotated, whereas the weakly supervised strategy just requires the annotation of the facial expressions. In order to combine the strengths of both fully supervised and weakly supervised regularization strategies, an hybrid formulation of them is also proposed.

The paper is organized in five sections including the Introduction (Section I). Section II presents the state-of-the-art methods for FER. The proposed neural network model along with the proposed regularization schemes are fully described in Section III. Section IV reports the experimental evaluation of the proposed methodology, in which a comparison with state-of-the-art methods is performed. Finally, conclusions and some topics for future work are presented in Section V.

## II. RELATED WORK

In the last decade, automatic facial expression recognition has been an active research topic in the artificial intelligence community due to its wide range of applications in the HCI field. Several facial expression recognition methodologies have been proposed, with an increasing progress in the recognition performance. An important part of this recent progress was achieved thanks to the emergence of deep learning approaches and more specifically with CNNs. Comprehensive surveys on automatic facial expression recognition can be found in [3] and [25]–[27].

Different deep architectures have been proposed for FER. Song *et al*. [28], developed a very simple FER system that uses a traditional CNN architecture composed of five layers. Some conventional training strategies, such as data augmentation and dropout, were applied in order to prevent overfitting. Similar approaches are proposed in [29]–[31]. In [29], a slightly more complex CNN architecture is presented. Inspired by the success of GoogleNet [32], the key structure of their architecture is a parallel feature extraction block that consists of convolutional, pooling, and ReLU layers. Tang *et al*. [33] reported a small but consistent advantage of replacing the softmax layer of the CNN with a linear support vector machine. The goal was to minimize a margin-based loss instead of the conventional cross-entropy loss function.

In most cases, a deep neural network model requires a lot of training data to generalize well, a condition that is not entirely fulfilled in the FER context where the amount of data is limited. Attempting to overcome this issue, several works have been using conventional deep learning regularization techniques (e.g., dropout, data augmentation, $l_2$-norm) along with *transfer learning* [34], [35], *classifier ensembles* [35]–[37], and *unsupervised learning* [38], which typically involves an unsupervised layer-wise training step that allows the usage of larger and unlabeled datasets.

Ng *et al*. [34] followed a transfer learning approach for deep CNN architectures, by utilizing a two-stage supervised fine-tuning process. More concretely, starting from a generic pre-training of two different CNN architectures based on the ImageNet dataset [39], a cascade fine-tuning approach is, then, applied using two different facial expression datasets. Yu and Zhang [35] propose a classification module that consists of an ensemble of multiple deep CNNs. Each CNN model is randomly initialized and pre-trained in a larger dataset before being fine-tuned on the target dataset. To combine multiple CNN models, they propose two constrained optimization frameworks to automatically learn the ensemble weights of the network responses. Similar approaches are proposed in [36] and [37]. The authors propose a hierarchical architecture of a committee of deep CNNs with an exponentially-weighted decision fusion. The individual CNN models were trained varying the network architecture, input normalization and weight initialization, in order to obtain diverse decisions boundaries.

More recently, Connie *et al*. [40] proposed an hybrid approach, in which SIFT features are merged with one of the later CNN layers. The underlying idea is to combine the strengths of hand-crafted and deep learning approaches. Experimental results suggest that the fusion approach yields an overall improvement in the FER performance.

Other deep learning techniques, such as deep belief networks (DBNs), have also been used for FER [38], [41]. It is the example of the work of Liu *et al*. [38], in which a two-step iterative learning process is used to train boosted DBNs. First, each DBN learns a non-linear feature representation from a facial patch in an unsupervised manner. Second, these DBNs are connected through a boosted classifier and fine-tuned

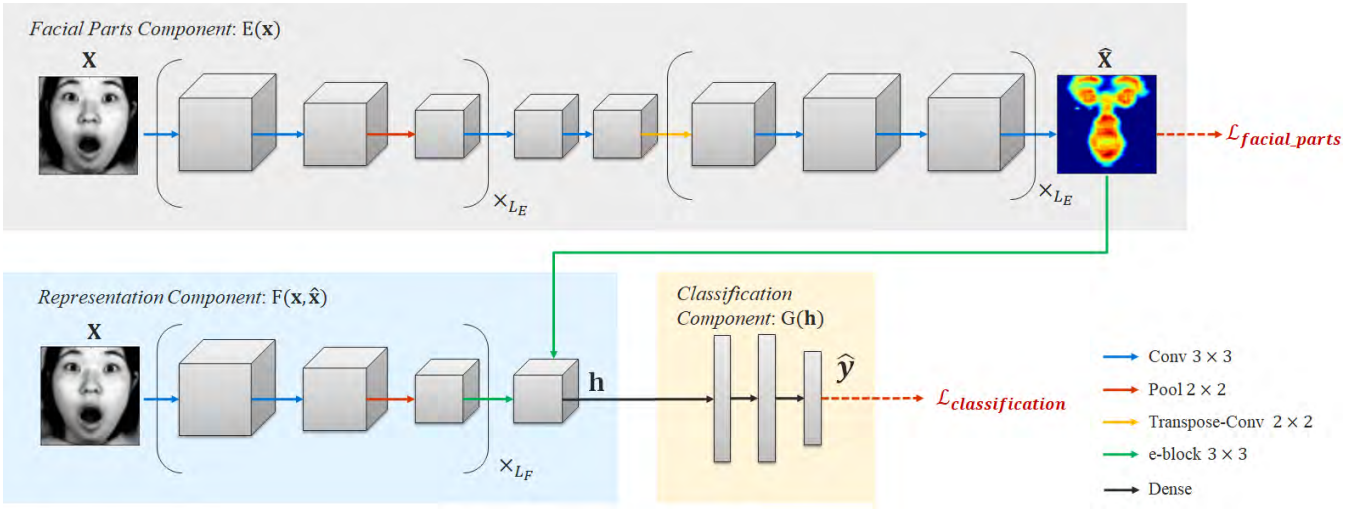



**FIGURE 4.** The architecture of the proposed neural network for FER. It comprises three modules or components, i.e. the *facial-parts component*, the *representation component*, and the *classification component*.

jointly driven by a single objective function. In this regard, the features extracted at different locations are selected and strengthened jointly according to their relative importance to the facial expression recognition. Liu *et al*. [41] propose the so-called AU-aware deep networks, in which a fixed convolutional step (i.e., application of a predefined set of hand-crafted filters) followed by a pooling step is applied to extract a feature representation. Then, the representation is grouped into a set of relevant receptive fields for each expression. Each receptive field is fed to a DBN to obtain a non-linear feature representation, using an SVM to detect each expression independently.

In terms of motivation, the work of Liu *et al*. [41] is probably the most related to our proposed methodology, as they also explore the psychological theory that FEs can be decomposed into multiple action units. However, it should be noticed that the proposed neural network architecture, objective function, as well as the entire learning strategy, are completely different. First, the neural network proposed in [41] is not trained end-to-end. Second, they do not explore the potential of CNNs to extract expression-specific representations. As they use a set of hand-crafted filters, the modeling capacity of their model is limited by the fixed transformations (filters).

## III. PROPOSED MODEL

While *transfer learning* across tasks has been widely applied to work around the challenge of training deep models in small datasets, such as those available for FER, the benefits of *transfer learning* are tightly coupled with the source-target domain similarity. Instead, our goal is to design a deep model by imposing domain knowledge based on the strong support from physiology and psychology that FEs are the result of the motions of facial muscles [3], [23]. The underlying idea is to explicitly drive the model towards the most relevant facial areas for the expression recognition, such as the facial components (i.e., eyes, eyebrows, nose, mouth) and expression wrinkles.

In this regard, we propose a novel deep neural network architecture along with a well-designed loss function that explicitly models both informative local facial regions and expression recognition. The result is a model that is able to jointly learn facial relevance maps and expression-specific features for a proper recognition.

To induce the model to jointly learn the most relevant facial parts along with the FER, the proposed neural network is composed by three main components, namely (i) the *facial-parts component*, (ii) the *representation component*, and (iii) the *classification component*. The purpose of the *facial-parts component* is to learn an encoding-decoding function $E(\mathrm{x};\theta_E)$, parameterized by $\theta_E$, that maps from an input image x to a relevance map $\hat{\mathrm{x}}$ representing the probability of each pixel being relevant for recognition. The loss function is defined in a such manner that enforces sparsity and spatial contiguity on the activations of $\hat{\mathrm{x}}$. This definition is supported by the physiological fact that just small and disjoint facial regions are relevant for recognition [23]. The *representation component* aims to learn an embedding function $F(\mathrm{x}, \hat{\mathrm{x}};\theta_F)$, parameterized by $\theta_F$, that maps from an input image x and its relevance map $\hat{\mathrm{x}}$ to an hidden representation h. The relevance map $\hat{\mathrm{x}}$ that is being learned in the *facial-parts component* is then used to filter the learned representations h, enforcing them to only respond strongly to the most relevant facial parts as possible. The result is a model that produces highly discriminative representations for FER. The *classification component* is then trained on these highly discriminative representations. Formally, $G(\mathrm{h};\theta_G)$ represents a task-specific function, parameterized by $\theta_G$, that maps from hidden representations h to the task-specific predictions $\hat{\mathrm{y}}$.

### A. ARCHITECTURE

As shown in Figure 4, the architecture of the proposed neural network comprises three main modules, i.e. the *facial-parts component*, the *representation component*, and the *classification component*.

#### 1) FACIAL-PARTS COMPONENT

The architecture of the *facial-parts component* consists of a convolutional path followed by a deconvolutional path, in a such way that it is possible to learn a mapping between an input image x to a relevance map $\hat{x}$, with the same resolution of the input.

The convolutional path follows the typical architecture of a fully convolutional network [42]. It comprises several sequences of two consecutive $3 \times 3$ convolutional layers, with rectified linear units (ReLUs) as non-linearities, followed by a $2 \times 2$ max-pooling operation for downsampling. The number of convolutional filters is doubled at each max-pooling operation.

Every step in the deconvolutional path comprises a $2 \times 2$ transpose convolution and two $3 \times 3$ convolutions, each one followed by a ReLU. The transpose convolution is applied for up-sampling and densify the incoming features maps. At the final layer a $3 \times 3$ convolution with a linear activation function is used to map the activations into a probability relevance map.

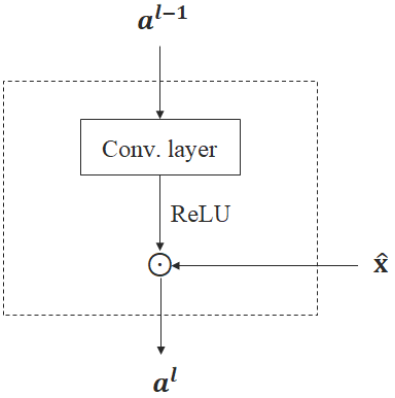



**FIGURE 5.** The expression block (e-block).

#### 2) REPRESENTATION COMPONENT

The purpose of the *representation component* is to extract highly discriminative features for FER. Therefore, it starts with several sequences of convolution-convolution-pooling layers for a typical CNN feature extraction. Then, we introduce a novel building block in the network, the so-called expression block (e-block), in order to increase the discriminative ability of the learned features. As illustrated in Figure 5, an e-block comprises a convolutional layer and a elementwise multiplication. It takes as input the activations of the previous layer (the learned features) and the relevance map $\hat{x}$. Formally, the e-block is defined as:

$$a^l = \sigma(W * a^{l-1}) \odot \hat{x}, \tag{1}$$

where $*$ and $\odot$ denote a convolution operation and a elementwise multiplication, respectively. $a^{l-1}$ and $a^l$ represent the input and output activations of the e-block, respectively. $W$ represents the weights of the convolutional layer to be learned and $\sigma$ is the non-linearity (i.e., ReLU). The biases are omitted for notation simplification. The elementwise multiplication with $\hat{x}$ is performed to enforce the output activations $a^l$ to just respond strongly to the most relevant facial parts. It should be noticed that $\hat{x}$ has to be resized and cropped accordingly to the actual feature map size for a proper elementwise multiplication.

#### 3) CLASSIFICATION COMPONENT

The architecture of the *classification component* simply consists of a sequence of fully connected layers (or dense layers). The last layer of the CNN is a softmax output layer, which contains the output probabilities for each class label. The output node that produces the largest probability is chosen as the overall classification.

### B. LEARNING

Inference in the proposed model is given by $\hat{x} = E(x)$ and $\hat{y} = G(h)$ where $\hat{x}$ is the relevance map of the facial parts, $\hat{y}$ is the task-specific prediction and $h = F(x, \hat{x})$. Therefore, the goal of training is to minimize the following loss function with respect to parameters $\Theta = \{\theta_E, \theta_F, \theta_G\}$:

$$\mathcal{L} = \mathcal{L}_{\text{classification}} + \lambda \, \mathcal{L}_{\text{facial\_parts}}, \tag{2}$$

where $\lambda \geq 0$ is the weight that controls the interaction of the loss terms. The classification loss, $\mathcal{L}_{classification}$, trains the model to predict the output labels and corresponds to the categorical cross-entropy defined by:

$$\mathcal{L}_{\text{classification}} = -\sum_{i=1}^{N} y_i^{\top} \log \hat{y}_i, \tag{3}$$

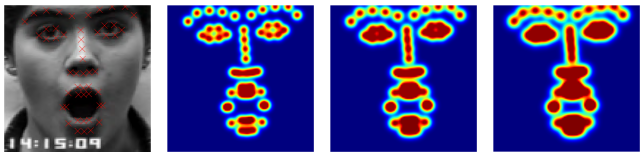
where $y_i$ is a column vector denoting the one-hot encoding of the class label for input $i$ and $\hat{y}_i$ are the softmax predictions of the model: $\hat{y}_i = G(h_i)$.

The purpose of the facial-parts loss, $\mathcal{L}_{facial\_parts}$, is to enforce the relevance map $\hat{x}$ to encode the relative importance of each pixel to the facial expression classification. Based on the physiological support that FEs can be decomposed into several action units of facial muscles, the underlying assumption is that the relevance map $\hat{x}$ should be sparse and spatially localized. It means that $\hat{x}$ should take high values just in the neighborhood of important facial components (e.g., eyes, eyebrows, nose, mouth and expression winkles).

To accomplish this purpose, we propose three different regularization strategies for regression of $\hat{x}$, accordingly to the level of the available data annotations:

#### 1) FULLY SUPERVISED REGULARIZATION

The proposed fully supervised regularization scheme requires not only the availability of the ground-truth class labels but also the annotation of the true coordinates of some



(a) (b) (c) (d)

**FIGURE 6.** Fully supervised learning scheme: (a) a training image with the true key-points coordinates superimposed (red crosses), and (b-d) examples of target relevance maps $\mathbf{x}^{target}$, obtained by a superposition of Gaussians at the location of each facial landmark, with an increasing $\sigma$ value.

facial landmarks (or key-points) located over important facial components, such as the eyes, nose, mouth and eyebrows (see Figure 6a).

In this scenario, a target relevance map $x_i^{target}$ for each training image $i$ is created, $i = 1, ..., N$. Let $\mathrm{K} = \{(r, c)^j\}_{i=1}^N$, $j = 1, ..., k$, represent the set all $k$ annotated key-points coordinates. As illustrated in Figure 6, for a given training image, each facial landmark $j$ is represented by a Gaussian, with mean at the key-point coordinates, i.e., $\mu = (r, c)^j$, and a predefined standard deviation $\sigma$. Then, the target relevance map $x_i^{target}$ is simply formed by the mixture of the Gaussians of each facial landmark. The standard deviation $\sigma$ should be set to control the neighborhood size around the facial landmarks (see Figures 6b-6d).

The facial-parts loss, $\mathcal{L}_{facial\_parts}$, is then defined to minimize the mean squared error between the target and the predicted relevance maps, such that:

$$\mathcal{L}_{\text{facial\_parts}} = \frac{1}{N}\sum_{i=1}^{N}(x_i^{target} - \hat{x}_i)^2 \tag{4}$$

Therefore, this loss term encourages the relevance map $\hat{\mathrm{x}}$ to take high values in the neighborhood of the most important facial components.

### 2) WEAKLY SUPERVISED REGULARIZATION

The weakly supervised regularization strategy does not require the annotation of the facial key-points coordinates. In this scenario, the facial-parts loss, $\mathcal{L}_{facial\_parts}$, is defined to regularize the activations of the relevance map $\hat{x}$ by imposing sparsity and spatial contiguity as follows:

$$\mathcal{L}_{\text{facial\_parts}} = \sum_{i=1}^{N}\mathcal{L}_{\text{sparsity}}(\hat{x}_i) + \gamma\sum_{i=1}^{N}\mathcal{L}_{\text{contiguity}}(\hat{x}_i), \tag{5}$$

where $\gamma \geq 0$ is the weight that controls the interaction of the loss terms. The intuition is that just small and disjoint facial regions are relevant for the recognition task. In this regard, the sparsity term is defined by:

$$\mathcal{L}_{\text{sparsity}}(\hat{x}) = \frac{1}{m \times n}\sum_{i,j}|\hat{x}_{i,j}|, \tag{6}$$

where $m$, $n$ denote the resolution of the relevance map $\hat{x}$.

The spatial contiguity term $\mathcal{L}_{\text{contiguity}}$ encourages the activations of $\hat{x}$ to be smooth and spatially localized. Then, the spatial contiguity loss is simply defined to minimize the local spatial transitions of the relevance map $\hat{x}$, as follows:

$$\mathcal{L}_{\text{contiguity}}(\hat{x}) = \frac{1}{m \times n}\sum_{i,j}|\hat{x}_{i+1,j} - \hat{x}_{i,j}| + |\hat{x}_{i,j+1} - \hat{x}_{i,j}| \tag{7}$$

It should be noticed that, as defined, $\mathcal{L}_{\text{sparsity}}$ and $\mathcal{L}_{\text{contiguity}}$ correspond to the $l_1$ regularization and the total variation regularization on the activations of $\hat{x}$, respectively. In fact, the $\mathcal{L}_{\text{sparsity}}$ term could have been defined as the $l_0$-norm, since the $l_0$-optimization has also the property of producing sparse solutions [43], [44]. However, the corresponding $l_0$-optimization problem is non-convex and, hence, difficult to solve. It is known to be NP-hard. In this regard, $\mathcal{L}_{\text{sparsity}}$ was defined as the $l_1$-norm, since $l_1$ is indeed a good differentiable approximation to $l_0$ [44].

### 3) HYBRID FULLY AND WEAKLY SUPERVISED REGULARIZATION

In a completely annotated scenario, i.e., when both expression labels and facial landmarks annotations are available, the regression task of the relevance map $\hat{x}$ can be performed by combining both fully and weakly supervised regularization schemes. In this case, the facial-parts loss to be minimized is simply defined as the weighted summation of the loss terms defined in Equations 4 and 5. The underlying idea is to combine the strengths of both proposed regularization schemes while mitigating their potential weaknesses when used individually.

The proposed fully supervised regularization scheme encourages the predicted relevance maps $\hat{\mathrm{x}}$ to be as similar as possible to the target ones $\mathbf{x}^{target}$. In this regard, the resulting relevance maps $\hat{\mathrm{x}}$ will "just" encode the local appearance information around the facial landmarks. Although the facial landmarks, along with the facial components in which they lay on, could represent some of the most relevant facial areas for expression recognition, other important facial clues, such as the expression wrinkles or dimples, may be neglected. As the weakly supervised regularization scheme relies on sparsity and contiguity impositions, the resulting relevance maps will have the potential to capture expressions wrinkles and dimples. However, as the relevance maps are learned with no supervision, the optimization process is more difficult and highly sensitive to the hyperparameters choice ($\lambda$ and $\gamma$).

By combining both regularization schemes, the predicted relevance maps $\hat{\mathrm{x}}$ will encode local appearance information around facial landmarks with the freedom to capture additional sparse and contiguity facial features, such as expression wrinkles and dimples.

## C. DATA AUGMENTATION

In general, data augmentation is the process of increasing, artificially, the number of training samples, by means of different image transformations and/or noise addition. In here, a randomized data augmentation scheme based on both

**FIGURE 7.** Illustration of the implemented data augmentation process: original colour images (top row) along with the corresponding augmented images (bottom row).

geometric and colour transformations is applied during the training step. The underlying idea is to increase the robustness of the proposed model to the wide range of face positions, poses, viewing angles as well as to different illumination conditions and contrasts. The data augmentation process is applied in an online-fashion, within every iteration, to all the images of each mini-batch.

Specifically, the considered geometric transformations are obtained through the following randomized affine image warping:

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} s & 0 \\ 0 & s \end{bmatrix} \begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix} \begin{bmatrix} 1 & k_1 \\ k_2 & 1 \end{bmatrix} \begin{bmatrix} x - t_1 \\ y - t_2 \end{bmatrix}, \tag{8}$$

where $\theta$ is the rotation angle, $k_1$ and $k_2$ are the skew parameters along the $x$ and $y$ directions. $t_1$ and $t_2$ denote both translation parameters and $s$ is the scale factor. It is important to note that the values of these parameters are randomly selected from predefined sets, as listed in Section IV. Pixels mapped outside the original image are assigned with the pixel values of their mirrored position.

The other type of image augmentation focuses on randomly normalizing the contrast of each channel in the training images. Formally, let $S_c$ be the $c$-th channel of the input image, the new intensity value at each pixel in channel $c$ is simply given by:

$$S'_c = \begin{cases} 0, & \text{if } S_c < S_c(p_L) \\ \dfrac{S_c - S_c(p_L)}{S_c(p_H) - S_c(p_L)}, & \text{if } S_c(p_L) \leq S_c \leq S_c(p_H) \\ 1, & \text{if } S_c > S_c(p_H), \end{cases} \tag{9}$$

where $p_L$ and $p_H$ represent the lower and higher histogram percentiles that are randomly selected for the colour transformation, respectively. This scheme simulates the scenario that the input images are acquired with different intensities, contrasts and illuminations conditions. Figure 8 illustrates the application of the implemented data augmentation procedure.

## IV. EXPERIMENTAL EVALUATION

The experimental evaluation of the proposed deep neural network was performed using public available databases in the FER research field: the Extended Cohn-Kanade (CK+) database [45], the Japanese Female Facial Expressions (JAFFE) database [46], the Static Facial Expressions in the Wild (SFEW) database [47] and the Facial Expression Recognition 2013 (FER-2013) database [48].

While both CK+ and JAFFE datasets contain images acquired under lab-controlled conditions, SFEW 2.0 and FER contain images with spontaneous expressions acquired under wild (non-controlled) scenarios. Table 1 depicts the total number of images of each dataset as well as the class distribution. Some representative samples of each dataset are shown in Figure 8.

**TABLE 1.** Summary of the datasets used in the experimental evaluation.

| | Neutral | Anger | Contempt | Disgust | Fear | Happy | Sadness | Surprise | Total |
|---|---|---|---|---|---|---|---|---|---|
| CK+ [45] | 327 | 135 | 54 | 177 | 75 | 147 | 84 | 249 | 1308 |
| JAFFE [46] | - | 30 | - | 41 | 8 | 54 | 39 | 41 | 213 |
| SFEW [47] | 228 | 255 | - | 75 | 124 | 256 | 234 | 150 | 1322 |
| FER-2013 [48] | 6198 | 4953 | - | 547 | 5121 | 8989 | 6077 | 4002 | 35887 |

### A. IMPLEMENTATION DETAILS

As a pre-processing step, the multi-task CNN face detector [49] is used for face detection. The faces are then normalized, cropped, and resized to $120 \times 120$ pixels.

The proposed fully supervised regularization scheme as well as the proposed hybrid regularization strategy require the facial landmarks annotation for creating the target relevance maps $\mathrm{x}^{target}$. Although the CK+ dataset contains the annotations of the facial landmarks, these manual annotations are not available on both JAFFE and SFEW. Therefore, a robust facial landmarks detector [50] was applied on both JAFFE and SFEW datasets and, then, the automatically generated facial landmarks were used to build the target relevance maps.

All deep models are implemented in Theano [51] and trained with the Adam optimization algorithm using a batch size of 50 samples. We used a learning rate with step decay, in which the initial learning rate was multiplied by 0.99 at each training epoch.

The hyperparameters of the models are optimized by means of grid search and cross-validation on the training set. These parameters include the weights of all loss terms ($\lambda$ and $\gamma$), the learning rate $\alpha$, the $l_2$ coefficient, and the number of convolution-convolution-pooling blocks of both *facial-parts* and *representation components* ($L_E$ and $L_F$, respectively). The number of dense layers of the *classification component* was set to 3 in all the experiments. In particular, while the number of neurons of the last dense layer (i.e., the output layer) corresponds to the number of classes, the first two dense layers contain 512 neurons. A detailed description of the architecture of the proposed model is presented in Table 3. For a fair comparison, the hyperparameters (i.e., architecture, learning rate and $l_2$ coefficient) of the CNN trained from scratch as baseline were also optimized. The range of values of the adopted hyperparameters' grid search is presented in Table 2.

Regarding the parameters of the data augmentation scheme, the rotation angle $\theta$ is randomly sampled from $\{-\pi/18, -\pi/36, 0, \pi/36, \pi/18\}$. The skew parameters,


**FIGURE 8.** Illustration of the implemented data augmentation process: original colour images (top row) along with the corresponding augmented images (bottom row). (a) CK + [45]. (b) JAFFE [46]. (c) SFEW 2.0 [47]. (d) FER-2013 [48].

**TABLE 2.** Hyperparameters sets.

| Hyperparameters | Acronym | Set |
|---|---|---|
| Architecture | $L_E$ | {3,4} |
| | $L_F$ | {3,4} |
| Leaning rate | $\alpha$ | $\{1e^{-03}, 1e^{-04}\}$ |
| $l_2$-norm coefficient | - | $\{1e^{-04}, 1e^{-05}\}$ |
| Facial parts loss† | $\lambda$ | {1,5,10,15} |
| Facial parts loss‡ | $\lambda$ | $\{1e^{-03}, 1e^{-04}, 1e^{-05}, 1e^{-06}\}$ |
| | $\gamma$ | $\{\frac{1e^{-03}}{\lambda}, \frac{1e^{-04}}{\lambda}, \frac{1e^{-05}}{\lambda}, \frac{1e^{-06}}{\lambda}\}$ |

† fully supervised regularization scheme, ‡ weakly supervised regularization scheme.

**TABLE 3.** A detailed description of the architecture of the proposed model. The output shape is described as (#filters, rows, columns).

| Layer # | Network module | Layer (type) | Output shape | Connected to |
|---|---|---|---|---|
| 1 | | input_1 (InputLayer) | (3, 120, 120) | - |
| 2 | | conv2d_1 (Conv2D) | (16, 120, 120) | input_1 |
| 3 | | conv2d_2 (Conv2D) | (16, 120, 120) | conv2d_1 |
| 4 | | max_pool2d_1 (MaxPooling2D) | (16, 60, 60) | conv2d_2 |
| 5 | | conv2d_3 (Conv2D) | (32, 60, 60) | max_pool2d_1 |
| 6 | | conv2d_4 (Conv2D) | (32, 60, 60) | conv2d_3 |
| 7 | | max_pool2d_2 (MaxPooling2D) | (32, 30, 30) | conv2d_4 |
| 8 | | conv2d_5 (Conv2D) | (64, 30, 30) | max_pool2d_2 |
| 9 | | conv2d_6 (Conv2D) | (64, 30, 30) | conv2d_5 |
| 10 | | max_pool2d_3 (MaxPooling2D) | (64, 15, 15) | conv2d_6 |
| 11 | | conv2d_7 (Conv2D) | (128, 15, 15) | max_pool2d_3 |
| 12 | | conv2d_8 (Conv2D) | (128, 15, 15) | conv2d_7 |
| 13 | $E(\mathrm{x})$ | conv2d_tr_1 (Conv2DTranspose) | (64, 30, 30) | conv2d_8 |
| 14 | | concat_1 (Concatenate) | (128, 30, 30) | [conv2d_tr_1; conv2d_6] |
| 15 | | conv2d_9 (Conv2D) | (64, 30, 30) | concat_1 |
| 16 | | conv2d_10 (Conv2D) | (64, 30, 30) | conv2d_9 |
| 17 | | conv2d_tr_2 (Conv2DTranspose) | (32, 60, 60) | conv2d_10 |
| 18 | | concat_2 (Concatenate) | (64, 60, 60) | [conv2d_tr_2; conv2d_4] |
| 19 | | conv2d_11 (Conv2D) | (32, 60, 60) | concat_2 |
| 20 | | conv2d_12(Conv2D) | (32, 60, 60) | conv2d_11 |
| 21 | | conv2d_tr_3 (Conv2DTranspose) | (16, 120, 120) | conv2d_12 |
| 22 | | concat_3 (Concatenate) | (32, 120, 120) | [conv2d_tr_3; conv2d_2] |
| 23 | | conv2d_13 (Conv2D) | (16, 120, 120) | concat_3 |
| 24 | | conv2d_14 (Conv2D) | (1, 120, 120) | conv2d_13 |
| 25 | | conv2d_15 (Conv2D) | (16, 120, 120) | input_1 |
| 26 | | conv2d_16 (Conv2D) | (16, 120, 120) | conv2d_15 |
| 27 | | max_pool2d_4 (MaxPooling2D) | (16, 60, 60) | conv2d_16 |
| 28 | | conv2d_17 (Conv2D) | (32, 60, 60) | max_pool2d_4 |
| 29 | | conv2d_18 (Conv2D) | (32, 60, 60) | conv2d_17 |
| 30 | $F(\mathrm{x}, \hat{\mathrm{x}})$ | max_pool2d_5 (MaxPooling2D) | (32, 30, 30) | conv2d_18 |
| 31 | | conv2d_19 (Conv2D) | (64, 30, 30) | max_pool2d_5 |
| 32 | | conv2d_20 (Conv2D) | (64, 30, 30) | conv2d_19 |
| 33 | | max_pool2d_6 (MaxPooling2D) | (64, 15, 15) | conv2d_20 |
| 34 | | eblock_1 (e-block) | (64, 15, 15) | [max_pool2d_6; conv2d_14] |
| 35 | | dense_1 (Dense) | (512) | eblock_1 |
| 36 | | dropout_1 (Dropout) | (512) | dense_1 |
| 37 | $G(\mathrm{h})$ | dense_2 (Dense) | (512) | dropout_1 |
| 38 | | dropout_2 (Dropout) | (512) | dense_2 |
| 39 | | dense_3 (Dense) | (8) | dropout_2 |

$k_1$ and $k_2$, are both randomly sampled from $\{-0.1, 0, 0.1\}$. The scale parameter $s$ is randomly sampled from five different resize factors {0.9, 0.95, 1, 1.05, 1.1}. Finally, the translation parameters $t_1$ and $t_2$ are randomly sampled integers from the interval [0, 5].

### B. RELEVANCE MAPS VISUALIZATION

In order to demonstrate the effectiveness of the proposed deep model in capturing high-level semantic concepts related to facial expressions, we have performed a visual inspection of the relevance maps $\hat{\mathrm{x}}$ that are learned by the *facial-parts component* of our model. Figures 9 and 10 depict the learned relevance maps $\hat{\mathrm{x}}$ for some test samples using the proposed fully supervised and weakly supervised regularization schemes, respectively. As expected, the activations of the predicted relevance maps using both training schemes are strong just in the neighborhood of important facial components. This demonstrates that the relevance maps are suitable to enforce the model to learn highly discriminative representations for FER.

The fully supervised regularization scheme minimizes the mean squared error between the predicted relevance maps $\hat{\mathrm{x}}$ and the targets $\mathrm{x}^{target}$, which are created based on the facial landmarks location. Therefore, the predicted relevance maps encode the local appearance information around the facial landmarks. The weakly supervised regularization scheme does not rely on the facial landmarks location. Instead, the activations of the predicted relevance maps are regularized

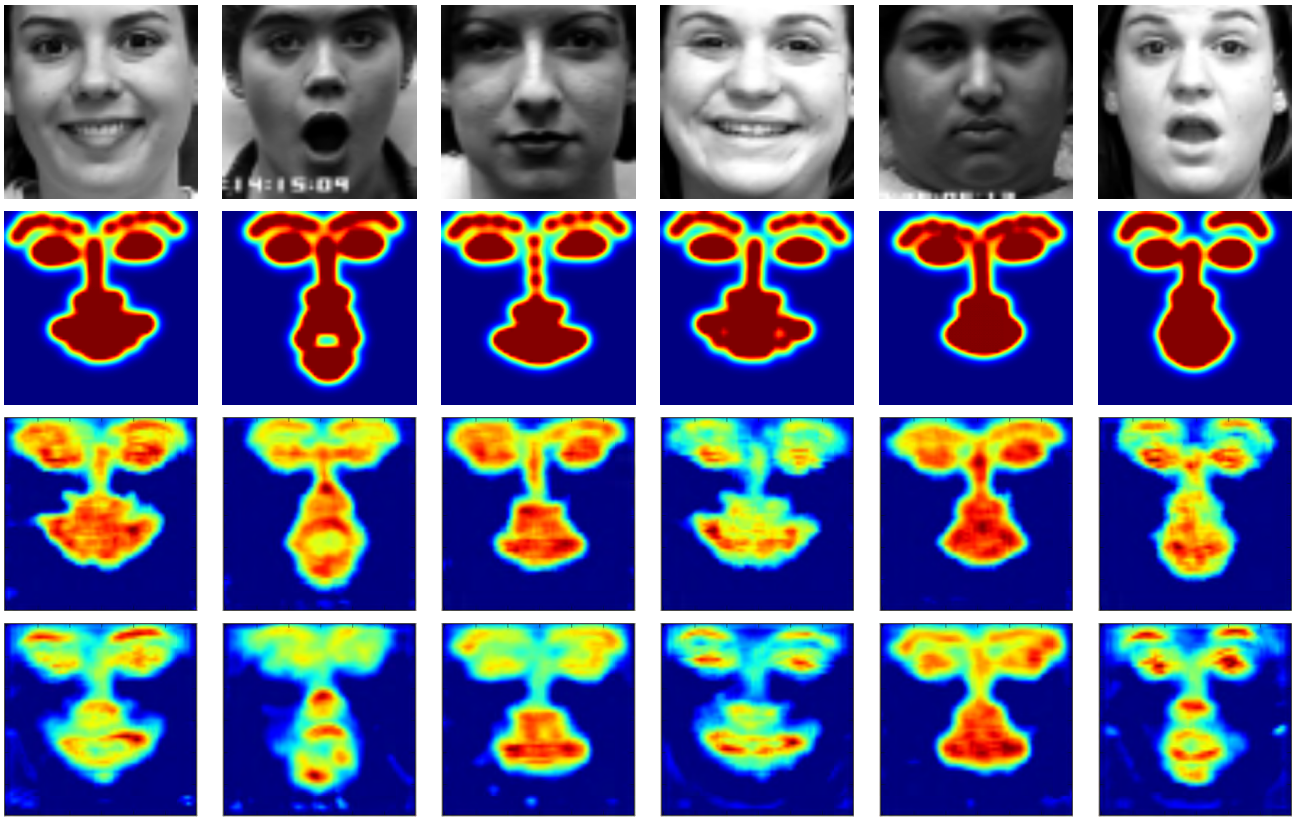


**FIGURE 9.** Illustrative examples of the predicted relevance maps $\hat{\mathbf{x}}$ using the proposed fully supervised regularization scheme and the effect of varying the facial-parts loss $\mathcal{L}_{facial\_parts}$ coefficient: the input images (first row) and the corresponding target relevance maps $\mathbf{x}^{target}$ (second row), predicted relevance maps $\hat{\mathbf{x}}$ with $\lambda = 10$ (third row), and predicted relevance maps $\hat{\mathbf{x}}$ with $\lambda = 5$ (bottom row).

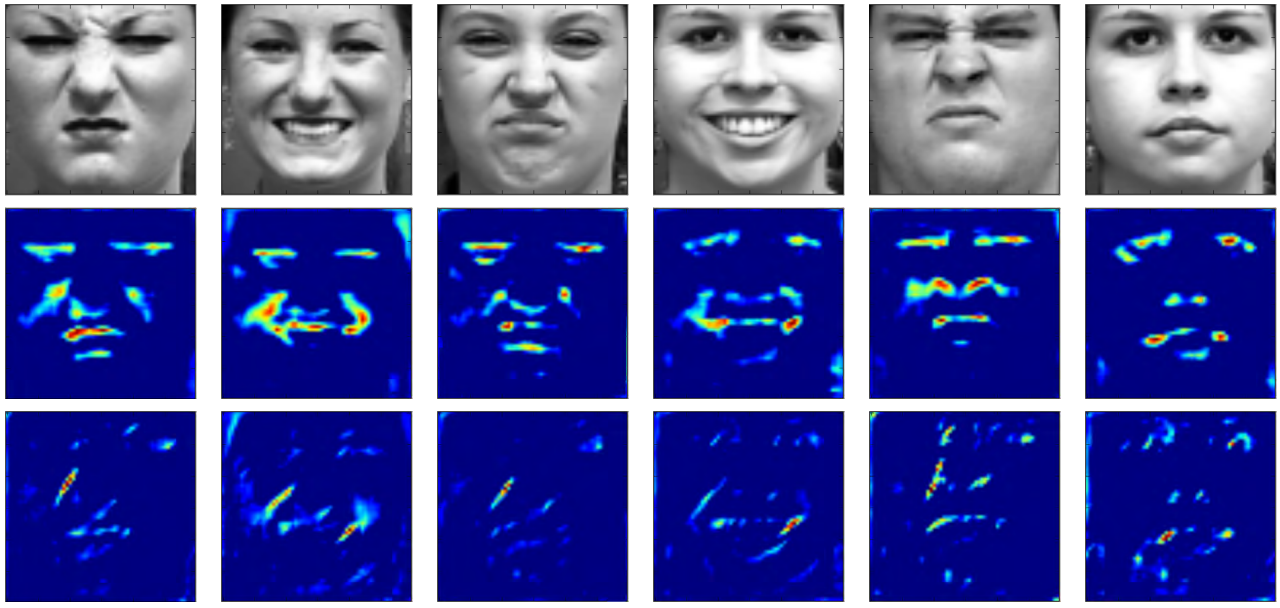


**FIGURE 10.** Illustrative examples of the predicted relevance maps $\hat{\mathbf{x}}$ using the proposed weakly supervised regularization scheme and the effect of varying the coefficients of $\mathcal{L}_{\text{sparsity}}$ and $\mathcal{L}_{\text{contiguity}}$: the input images (first row) and the corresponding predicted relevance maps $\hat{\mathbf{x}}$ with $\lambda = 1\text{e}^{-04}$ and $\gamma = 1$ (middle row), and predicted relevance maps with $\lambda = 1\text{e}^{-02}$ and $\gamma = 1$ (bottom row).

to be sparse and spatially localized. Interestingly, the resulting relevance maps are able to capture not only the local information around the facial landmarks but also the local information related to expression winkles (see the middle row of Figure 10). This clearly demonstrates the importance of the expression winkles to the recognition process.

Figures 9 and 10 also demonstrate the effect of varying the coefficients of the facial-parts loss $\mathcal{L}_{facial\_parts}$. Regarding the fully supervised version of the proposed model, as we decrease the $\lambda$ coefficient, the predicted relevance maps $\hat{x}$ are allowed to be more distant from the targets $x^{target}$ (see the bottom row of Figure 9). In the weakly supervised setting, as we increase the coefficients of the sparsity and the spatial contiguity terms ($\lambda$ and $\gamma$), the activations of the predicted relevance maps $\hat{x}$ are forced to be sparser and smoother (i.e., with less transitions), respectively. The middle row of Figure 10 illustrates the effect of setting a good parameterization to the coefficients of $\mathcal{L}_{\text{sparsity}}$ and $\mathcal{L}_{\text{contiguity}}$ (i.e., $\lambda = 1e^{-04}$ and $\gamma = 1$), which results in well defined relevance maps around the facial components (e.g., mouth, eyes, nose and expression winkles). The effect of an over-regularization is depicted in the bottom row of Figure 10 (i.e., $\lambda = 1e^{-02}$ and $\gamma = 1$). The resulting relevance maps are then too sparse and not so well defined around the facial components.

### C. RESULTS ON CK+

CK+ consists of 593 videos from 123 subjects acquired in a controlled environment, 327 of them annotated with 8 expression labels (i.e., the 6 universal expressions plus the neutral and contempt ones). Each video starts with a neutral expression and reaches the peak in the last frame. As in other works [38], the first frame and the last three frames of each video were extracted, in order to construct our image-based CK+ dataset. The result is a subset of 1308 images. For model selection and evaluation, a stratified $k$-fold cross-validation scheme with subject independence was adopted (i.e., $k = 10$). In each split, the training set is further divided, also with subject independence, in 80% for training and 20% for validation.

**TABLE 4. CK+ experimental results. The results are reported in terms of average classification accuracy. The first block of the table presents the results of state-of-the-art methods. The second block depicts the results of all versions of the proposed model and the baseline CNN. Bold number indicates the best method with the highest average classification accuracy.**

| Method | Average Accuracy (%) |
|---|---|
| Liu *et al.* (2013) [24] | 92.10 |
| Ding *et al.* (2017) [52] | 88.70 |
| Ding *et al.* (2017) [52] | 89.90 |
| Ng *et al.* (2015) [34] | 93.20 |
| CNN from Scratch with Reg (baseline) | 90.48 |
| Fully Supervised | 92.54 |
| Weakly Supervised ($\mathcal{L}_{sparsity}$) | 93.26 |
| Weakly Supervised ($\mathcal{L}_{contiguity}$) | 91.70 |
| Weakly Supervised ($\mathcal{L}_{sparsity} + \mathcal{L}_{contiguity}$) | 93.37 |
| Hybrid Fully and Weakly Supervised | **93.64** |

Experiments on CK+ database are presented in Table 4, in which a comparison between the proposed model and state-of-the-art methods, including both traditional and deep learning-based approaches, is performed. The results are presented in terms of average classification accuracy. It is important to note that we just considered state-of-the-art methods that followed the same evaluation protocol (i.e., 1308 images with 8 expressions). To further demonstrate the effectiveness of the proposed method, a CNN trained from scratch with $l_2$ regularization was considered as baseline. The baseline CNN has the same network architecture as the *representation component* of the proposed model. As shown in Table 4, the implemented CNN is a fairly strong baseline, with an overall classification accuracy of 90.48%.

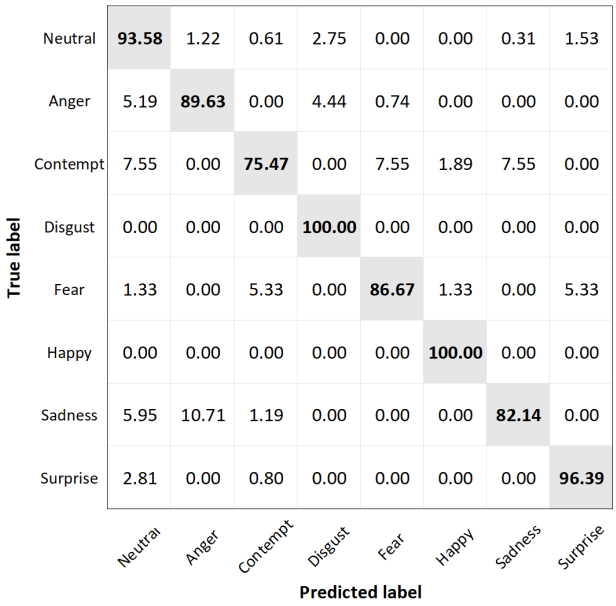



**FIGURE 11. Confusion matrix of CK+ dataset. Gray cells represent the true positives.**

Table 4 also depicts the performance of both versions of the proposed model (i.e., the fully and weakly supervised models) as well as their hybrid formulation (fully + weakly supervised). In order to assess the impact of the loss terms ($\mathcal{L}_{\text{sparsity}}$ and $\mathcal{L}_{\text{contiguity}}$) in the weakly supervised model, we report the results using each loss term independently and combined. Regardless the training strategy, the proposed model always outperforms the baseline CNN. In particular, the proposed hybrid model, which combines both fully supervised and weakly supervised regularization schemes, provides the best classification accuracy (93.64%), outperforming all the state-of-the-art methods.

One of the most interesting observations is the superior performance of the weakly supervised model when compared with the fully supervised model, despite the fact the weakly supervised model does not require the availability of the facial landmarks annotations. These results can be explained by the capability of the weakly supervised model in capturing local information around the expressions winkles (see the middle row of Figure 10). Another interesting observation, as also reported in Table 4, is that the proposed hybrid model, which combines the ideas of both regularization schemes, yields a slight overall improvement in the classification accuracy.

The confusion matrix, as illustrated in Figure 11, shows the consistent performance of the proposed hybrid method.

Both happy and disgust expressions are perfectly classified, while contempt is the most difficult to classify. This happens because the contempt expression is the class with the least number of training images and is typically performed in a subtle way.

### D. RESULTS ON JAFFE

The database contains 213 images of 6 facial expressions posed by 10 Japanese female models. Illustrative examples of the JAFFE dataset are shown in Figure 8b. For model selection and evaluation, a stratified 3-fold cross-validation with subject independence was performed. In each split, the training set is further divided, also with subject independence, in 80% for training and 20% for validation.

**TABLE 5. JAFFE experimental results. The results are reported in terms of average classification accuracy. The first block of the table presents the results of state-of-the-art methods. The second block depicts the results of all versions of the proposed model and the baseline CNN. Bold number indicates the best method with the highest average classification accuracy.**

| Method | Average Accuracy (%) |
|---|---|
| Shan *et al.* (2009) [12] | 81.00 |
| Lopes *et al.* (2017) [53] | 84.48 |
| Happy *et al.* (2015) [54] | 85.06 |
| CNN from Scratch with Reg (baseline) | 79.06 |
| Fully Supervised | 84.88 |
| Weakly Supervised ($\mathcal{L}_{sparsity}$) | 87.21 |
| Weakly Supervised ($\mathcal{L}_{contiguity}$) | 80.81 |
| Weakly Supervised ($\mathcal{L}_{sparsity} + \mathcal{L}_{contiguity}$) | 87.84 |
| Hybrid Fully and Weakly Supervised | **89.01** |

Table 5 compares the performance of the proposed approach with the baseline CNN and state-of-the-art methods. As observed from Table 5, the proposed fully supervised and weakly supervised models along with their hybrid version clearly outperform the implemented baseline CNN, with classification accuracies of 84.88%, 87.84%, 89.01% and 79.06%, respectively. In addition, our method provides substantial improvements over the previous best state-of-the-art performance achieved by Lopes *et al.* [54], with a gain of 3.95%. These results clearly demonstrate the potential of the proposed approach to deal with the problem of training high-capacity classifiers in small datasets (e.g., JAFFE database is composed by only 213 images).

Figure 12 shows the confusion matrix obtained for the best model on JAFFE, which is the proposed hybrid model. As it is possible to observe, the fear expression is perfectly classified. The proposed model performed worst for the surprise expression as it tends to be misclassified as happy.

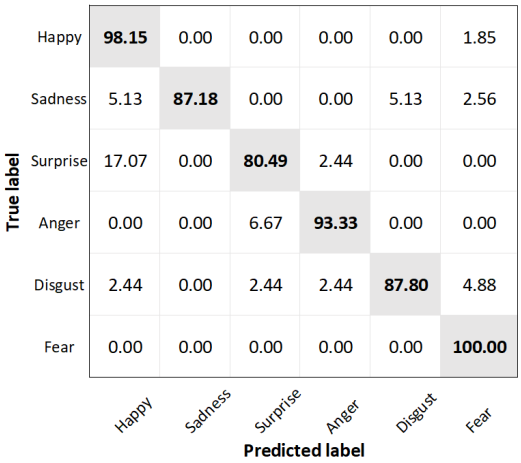



**FIGURE 12. Confusion Matrix of JAFFE dataset. Gray cells represent the true positives.**

### E. RESULTS ON SFEW

Different from both CK+ and JAFFE datasets, SFEW is targeted for unconstrained FER. It is the first database that depicts real-world or simulated real-world conditions for expression recognition. The images are all extracted from movies (see Figure 8c), and labeled with seven expressions. Therefore, there is a wide range of poses, viewing angles, occlusions, illumination conditions and, hence, the recognition is much more challenging. As SFEW was created as part of the Emotion Recognition in the Wild (EmotiW) 2015 Grand Challenge [55], it has a strict evaluation protocol with predefined training, validation, and test sets. In particular, the training set comprises a total of 891 images. Since we do not have access to the test set labels, the results are reported on the validation data that contains 431 images.

As SFEW is clearly one of the most challenging FER datasets, the top state-of-the-art methods on SFEW usually use other databases as additional training data. Typically, the current state-of-the-art models are pre-trained on FER-2013 before being fine-tuned to the SFEW dataset. The FER-2013 dataset comprises a total of 35887 grayscale images, labeled with seven facial expressions. The faces have been automatically registered so that the face is more or less centered and occupies about the same amount of space in each image (see Figure 8d). In this regard, for a fair comparison, the proposed model as well as the implemented baseline CNN are first pre-trained on the FER-2013 dataset and, then, fine-tuned to the target dataset (i.e., the SFEW). The fine-tunning process ends when the validation loss stops decreasing (∼25 epochs).

The experimental results obtained on SFEW are presented in Table 6, in which the state-of-the-art methods are grouped into those that do not perform any kind of transfer learning and those that use the FER-2013 dataset for pre-training the models. Once again, the proposed network clearly outperforms the implemented baseline CNN, with an overall accuracy of 50.12% against 42.07%. Moreover, the proposed method achieves better recognition rates than all the other state-of-the-art methods with the exception of the method proposed by Yu and Zhang [35]. However, we argue that this could not be a fair comparison as the method proposed in [35] uses an ensemble of multiple networks to boost their performance. In order to mitigate the gains of their ensemble strategy, a version of the Yu and Zhang [35] method without ensemble was implemented. As reported in Table 6, our

**TABLE 6. SFEW experimental results. The results are reported in terms of average classification accuracy. The first block of the table presents the results of state-of-the-art methods that do not use transfer learning. The second block of the table presents the results of state-of-the-art methods that use FER-2013 for pre-training the models. The third block depicts the results of all versions of the proposed model and the baseline CNN. Bold number indicates the best method with the highest average classification accuracy.**

| Method | Average Accuracy (%) | Transfer Learning |
|---|---|---|
| Liu *et al.* (2013) [24] | 26.14 | None |
| Liu *et al.* (2014) [38] | 31.73 | |
| Levi *et al.* (2015) [56] | 41.92 | |
| Mollahosseini *et al.* (2016) [57] | 47.70 | |
| Ng *et al.* (2015) [34] | 48.50 | FER-2013 |
| Yu *et al.* (2015) [35] | **52.29** | |
| Yu *et al.* (2015) [35] without ensemble † | 44.37 | |
| CNN from Scratch with Reg (baseline) | 42.07 | FER-2013 |
| Fully Supervised | 47.56 | |
| Weakly Supervised ($\mathcal{L}_{sparsity}$) | 48.72 | |
| Weakly Supervised ($\mathcal{L}_{contiguity}$) | 47.56 | |
| Weakly Supervised ($\mathcal{L}_{sparsity} + \mathcal{L}_{contiguity}$) | 47.80 | |
| Hybrid Fully and Weakly Supervised | 50.12 | |

† Implemented version of Yu *et al.* [35] method without ensemble.

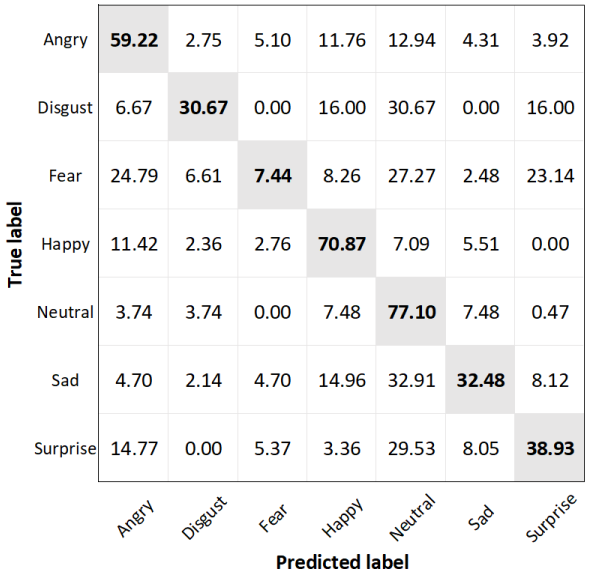



**FIGURE 13. Confusion Matrix of SFEW dataset. Gray cells represent the true positives.**

method clearly outperforms the method of Yu and Zhang [35] without ensemble (i.e, 50.12% against 44.37%).

Figure 13 shows the confusion matrix obtained for the proposed model with the best performance on SFEW (i.e., the hybrid fully and weakly supervised model). The recognition accuracy for fear is much lower than other expressions. This is also observed in other works [34].

## V. CONCLUSION

This paper addresses the topic of facial expression recognition on static images. In this regard, we propose a novel end-to-end deep neural network architecture along with a well-designed loss function that jointly learn the most relevant facial parts along with the expression recognition. The result is a model that is able to learn expression-specific features. The proposed neural network is composed by three main components: (i) the *facial-parts component*, (ii) the *representation component* and (iii) the *classification component*. The *facial-parts component* aims to regress a relevance map, representing the most important facial regions for the expression recognition. The relevance map is then used in the *representation component* in order to increase the discriminative ability of the learned features. Then, the *classification component* is trained on these highly discriminative representations for FER. Experimental results on three well-known facial expression databases CK+, JAFFE, and SFEW demonstrate the potential of the proposed model in both lab-controlled and wild scenarios. The proposed model provides quite promising results, outperforming in most datasets the current state-of-the-art methods. As future work, it is expected to apply the proposed network along with the training strategies to other domains with small datasets.

## REFERENCES

[1] S. Z. Li and A. K. Jain, *Handbook of Face Recognition*, 2nd ed. London, U.K.: Springer-Verlag, 2011.

[2] C. Darwin, *The Expression of the Emotions in Man and Animals*. London, U.K.: John Murray, 1982.

[3] C. A. Corneanu, M. O. Simón, J. F. Cohn, and S. E. Guerrero, "Survey on RGB, 3D, thermal, and multimodal approaches for facial expression recognition: History, trends, and affect-related applications," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 8, pp. 1548–1568, Aug. 2016.

[4] L. Zhang, D. Tjondronegoro, and V. Chandran, "Representation of facial expression categories in continuous arousal–valence space: Feature and correlation," *Image Vis. Comput.*, vol. 32, no. 12, pp. 1067–1079, 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0262885614001449

[5] R. Kosti, J. M. Alvarez, A. Recasens, and A. Lapedriza, "Emotion recognition in context," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1960–1968.

[6] S.-J. Wang *et al.*, "Micro-expression recognition using color spaces," *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 6034–6047, Dec. 2015.

[7] Z. Zhang, M. Lyons, M. Schuster, and S. Akamatsu, "Comparison between geometry-based and Gabor-wavelets-based facial expression recognition using multi-layer perceptron," in *Proc. 3rd IEEE Int. Conf. Autom. Face Gesture Recognit.*, Apr. 1998, pp. 454–459.

[8] I. Kotsia and I. Pitas, "Facial expression recognition in image sequences using geometric deformation features and support vector machines," *IEEE Trans. Image Process.*, vol. 16, no. 1, pp. 172–187, Jan. 2007.

[9] H. Tang and T. S. Huang, "3D facial expression recognition based on properties of line segments connecting facial feature points," in *Proc. 8th IEEE Int. Conf. Autom. Face Gesture Recognit.*, Sep. 2008, pp. 1–6.

[10] A. Maalej, B. B. Amor, M. Daoudi, A. Srivastava, and S. Berretti, "Shape analysis of local facial patches for 3D facial expression recognition," *Pattern Recognit.*, vol. 44, no. 8, pp. 1581–1589, 2011. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0031320311000756

[11] R. Walecki, O. Rudovic, V. Pavlovic, and M. Pantic, "Variable-state latent conditional random fields for facial expression recognition and action unit detection," in *Proc. 11th IEEE Int. Conf. Workshops Autom. Face Gesture Recognit. (FG)*, vol. 1, May 2015, pp. 1–8.

[12] C. Shan, S. Gong, and P. W. McOwan, "Facial expression recognition based on local binary patterns: A comprehensive study," *Image Vis. Comput.*, vol. 27, no. 6, pp. 803–816, 2009. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0262885608001844

[13] C. Liu and H. Wechsler, "Gabor feature based classification using the enhanced Fisher linear discriminant model for face recognition," *IEEE Trans. Image Process.*, vol. 11, no. 4, pp. 467–476, Apr. 2002.

[14] G. Littlewort *et al.*, "The computer expression recognition toolbox (CERT)," in *Proc. Face Gesture*, Mar. 2011, pp. 298–305.

[15] X. Sun, H. Xu, C. Zhao, and J. Yang, "Facial expression recognition based on histogram sequence of local Gabor binary patterns," in *Proc. IEEE Conf. Cybern. Intell. Syst.*, Sep. 2008, pp. 158–163.

[16] T. R. Almaev and M. F. Valstar, "Local Gabor binary patterns from three orthogonal planes for automatic facial expression recognition," in *Proc. Humaine Assoc. Conf. Affect. Comput. Intell. Interact.*, Sep. 2013, pp. 356–361.

[17] Z. Wang and Z. Ying, "Facial expression recognition based on local phase quantization and sparse representation," in *Proc. 8th Int. Conf. Natural Comput. (ICNC)*, May 2012, pp. 222–225.
[18] A. Dhall, A. Asthana, R. Goecke, and T. Gedeon, "Emotion recognition using PHOG and LPQ features," in *Proc. Face Gesture*, Mar. 2011, pp. 878–883.
[19] S. Berretti, A. Del Bimbo, P. Pala, B. B. Amor, and M. Daoudi, "A set of selected SIFT features for 3D facial expression recognition," in *Proc. 20th Int. Conf. Pattern Recognit.*, Aug. 2010, pp. 4125–4128.
[20] U. Tariq *et al.*, "Emotion recognition from an ensemble of features," in *Proc. Face Gesture*, Mar. 2011, pp. 872–877.
[21] A. Ramirez Rivera, R. Castillo, and O. Chae, "Local directional number pattern for face analysis: Face and expression recognition," *IEEE Trans. Image Process.*, vol. 22, no. 5, pp. 1740–1752, May 2013.
[22] Y. Rahulamathavan, R. C.-W. Phan, J. A. Chambers, and D. J. Parish, "Facial expression recognition in the encrypted domain based on local Fisher discriminant analysis," *IEEE Trans. Affective Comput.*, vol. 4, no. 1, pp. 83–92, Jan. 2013.
[23] P. Ekman and W. Friesen, *Facial Action Coding System: A Technique for the Measurement of Facial Movement*. Palo Alto, CA, USA: Consulting Psychologists Press, 1978
[24] M. Liu, S. Li, S. Shan, and X. Chen, "Au-aware deep networks for facial expression recognition," in *Proc. 10th IEEE Int. Conf. Workshops Autom. Face Gesture Recognit. (FG)*, Apr. 2013, pp. 1–6.
[25] G. Sandbach, S. Zafeiriou, M. Pantic, and L. Yin, "Static and dynamic 3D facial expression recognition: A comprehensive survey," *Image Vis. Comput.*, vol. 30, no. 10, pp. 683–697, Oct. 2012. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0262885612000935
[26] C.-D. Căleanu, "Face expression recognition: A brief overview of the last decade," in *Proc. IEEE 8th Int. Symp. Appl. Comput. Intell. Inform. (SACI)*, May 2013, pp. 157–161.
[27] E. Sariyanidi, H. Gunes, and A. Cavallaro, "Automatic analysis of facial affect: A survey of registration, representation, and recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 6, pp. 1113–1133, Jun. 2015.
[28] I. Song, H.-J. Kim, and P. B. Jeon, "Deep learning for real-time robust facial expression recognition on a smartphone," in *Proc. IEEE Int. Conf. Consum. Electron. (ICCE)*, Jan. 2014, pp. 564–567.
[29] P. Burkert, F. Trier, M. Z. Afzal, A. Dengel, and M. Liwicki, "DeXpression: Deep convolutional neural network for expression recognition," *CoRR*, vol. abs/1509.05371, pp. 1–8, Sep. 2015. [Online]. Available: http://arxiv.org/abs/1509.05371
[30] A. Uçar, "Deep convolutional neural networks for facial expression recognition," in *Proc. IEEE Int. Conf. Innov. Intell. Syst. Appl. (INISTA)*, Jul. 2017, pp. 371–375.
[31] K. Shan, J. Guo, W. You, D. Lu, and R. Bie, "Automatic facial expression recognition based on a deep convolutional-neural-network structure," in *Proc. IEEE 15th Int. Conf. Softw. Eng. Res., Manage. Appl. (SERA)*, Jun. 2017, pp. 123–128.
[32] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. IEEE Conf. CVPR*, Jun. 2015, pp. 1–9.
[33] Y. Tang, "Deep learning using linear support vector machines," *CoRR*, Jun. 2013. [Online]. Available: http://arxiv.org/abs/1306.0239
[34] H.-W. Ng, V. D. Nguyen, V. Vonikakis, and S. Winkler, "Deep learning for emotion recognition on small datasets using transfer learning," in *Proc. ACM Int. Conf. Multimodal Interact. (ICMI)*, New York, NY, USA, 2015, pp. 443–449, doi: 10.1145/2818346.2830593.
[35] Z. Yu and C. Zhang, "Image based static facial expression recognition with multiple deep network learning," in *Proc. ACM Int. Conf. Multimodal Interact. (ICMI)*, New York, NY, USA, 2015, pp. 435–442, doi: 10.1145/2818346.2830595.
[36] B.-K. Kim, H. Lee, J. Roh, and S.-Y. Lee, "Hierarchical committee of deep CNNs with exponentially-weighted decision fusion for static facial expression recognition," in *Proc. ACM Int. Conf. Multimodal Interact. (ICMI)*, New York, NY, USA, 2015, pp. 427–434, doi: 10.1145/2818346.2830590.
[37] B.-K. Kim, J. Roh, S.-Y. Dong, and S.-Y. Lee, "Hierarchical committee of deep convolutional neural networks for robust facial expression recognition," *J. Multimodal User Interfaces*, vol. 10, no. 2, pp. 173–189, Jun. 2016, doi: 10.1007/s12193-015-0209-0.
[38] M. Liu, S. Shan, R. Wang, and X. Chen, "Learning expressionlets on spatio-temporal manifold for dynamic facial expression recognition," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1749–1756.
[39] O. Russakovsky *et al.*, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.
[40] T. Connie, M. Al-Shabi, W. P. Cheah, and M. Goh, *Facial Expression Recognition Using a Hybrid CNN–SIFT Aggregator*. Cham, Switzerland: Springer, 2017, pp. 139–149.
[41] M. Liu, S. Li, S. Shan, and X. Chen, "AU-inspired deep networks for facial expression feature learning," *Neurocomput.*, vol. 159, pp. 126–136, Jul. 2015, doi: 10.1016/j.neucom.2015.02.011.
[42] S. Srinivas, R. K. Sarvadevabhatla, K. R. Mopuri, N. Prabhu, S. S. S. Kruthiventi, and R. V. Babu, "A taxonomy of deep convolutional neural nets for computer vision," *Frontiers Robot. AI*, vol. 2, p. 36, Jan. 2016. [Online]. Available: https://www.frontiersin.org/article/10.3389/frobt.2015.00036
[43] L. Xu, C. Lu, Y. Xu, and J. Jia, "Image smoothing via $L_0$ gradient minimization," *ACM Trans. Graph.*, vol. 30, no. 6, p. 174:1–174:12, Dec. 2011, doi: 10.1145/2070781.2024208.
[44] C. Ramirez, V. Kreinovich, and M. Argaez, "Why $\ell_1$ is a good approximation to $\ell_0$: A geometric explanation," *J. Uncertain Syst.*, vol. 7, no. 3, pp. 203–207, 2013.
[45] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended Cohn-Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2010, pp. 94–101.
[46] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba, "Coding facial expressions with Gabor wavelets," in *Proc. 3rd IEEE Int. Conf. Autom. Face Gesture Recognit.*, Nara, Japan, Apr. 1998, pp. 200–205.
[47] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon, "Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCV Workshops)*, Nov. 2011, pp. 2106–2112.
[48] I. J. Goodfellow *et al.* (2013). "Challenges in representation learning: A report on three machine learning contests." [Online]. Available: https://arxiv.org/abs/1307.0414
[49] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Process. Lett.*, vol. 23, no. 10, pp. 1499–1503, Oct. 2016.
[50] A. Bulat and G. Tzimiropoulos, "How far are we from solving the 2D & 3D face alignment problem? (and a dataset of 230,000 3D facial landmarks)," in *Proc. Int. Conf. Comput. Vis.*, 2017, pp. 1021–1030.
[51] R. Al-Rfou *et al.* (May 2016). "Theano: A Python framework for fast computation of mathematical expressions." [Online]. Available: https://arxiv.org/abs/1605.02688
[52] H. Ding, S. K. Zhou, and R. Chellappa, "FaceNet2ExpNet: Regularizing a deep face recognition net for expression recognition," in *Proc. 12th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2017, pp. 118–126.
[53] A. T. Lopes, E. de Aguiar, A. F. De Souza, and T. Oliveira-Santos, "Facial expression recognition with convolutional neural networks: Coping with few data and the training sample order," *Pattern Recognit.*, vol. 61, pp. 610–628, Jan. 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0031320316301753
[54] S. L. Happy and A. Routray, "Automatic facial expression recognition using features of salient facial patches," *IEEE Trans. Affective Comput.*, vol. 6, no. 1, pp. 1–12, Jan. 2015.
[55] (2015). *The Third Emotion Recognition in the Wild Challenge (EmotiW 2015)*. [Online]. Available: https://cs.anu.edu.au/few/emotiw2015.html
[56] G. Levi and T. Hassner, "Emotion recognition in the wild via convolutional neural networks and mapped binary patterns," in *Proc. ACM Int. Conf. Multimodal Interact. (ICMI)*, New York, NY, USA, 2015, pp. 503–510, doi: 10.1145/2818346.2830587.
[57] A. Mollahosseini, D. Chan, and M. H. Mahoor, "Going deeper in facial expression recognition using deep neural networks," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2016, pp. 1–10.

**PEDRO M. FERREIRA** received the degree in biomedical engineering from the Politécnico do Porto in 2009, and the M.Sc. degree in biomedical engineering from the Faculdade de Engenharia da Universidade do Porto (FEUP) in 2012. He is currently pursuing the Ph.D. degree enrolled in the Doctoral Program in electrical and computer engineering at FEUP. He is a Researcher at INESC TEC. His main research interests include Computer Vision, Machine Learning and Artificial Intelligence.

**FILIPE MARQUES** was born in Porto, Portugal, in 1995. He received the M.Sc. degree in bioengineering from the Faculdade de Engenharia da Universidade do Porto (FEUP) in 2018. He enrolled a curricular Research Internship at the Erasmus Medical Center, Biomedical Imaging Group, Rotterdam, The Netherlands, in 2017, focused on computer vision. During his master's project, he did a Research Internship at INESC TEC. His main interests are computer vision, machine learning, and artificial intelligence.

**JAIME S. CARDOSO** received the Licenciatura (5-year degree) in electrical and computer engineering in 1999, the M.Sc. degree in mathematical engineering in 2005, and the Ph.D. degree in computer vision in 2006, all from the University of Porto. He is currently an Associate Professor with Habilitation at the Faculty of Engineering of the University of Porto (FEUP) and also a Co-ordinator of the Centre for Telecommunications and Multimedia, INESC TEC. He has co-authored more than 200 papers, over 60 of which in international journals, which attracted over 2800 citations, according to Google scholar. His research can be summed up in three major topics: computer vision, machine learning, and decision support systems. Image and video processing focuses on biometrics and video object tracking for applications such as surveillance and sports. The work on machine learning cares mostly with the adaptation of learning to the challenging conditions presented by visual data. The particular emphasis of the work in decision support systems goes to medical applications, always anchored on the automatic analysis of visual data.

He was the Principal Investigator of eight research projects and has participated in 15 other research projects, including five European projects and a direct contract with BBC, the UK TV broadcaster. The know-how acquired in these research projects contributed to the creation of the company ClusterMedia Labs in 2006, for which he is the Co-founder.

**ANA REBELO** was born in Porto, Portugal, in 1985. She received the degree in mathematics applied to technology from the School of Sciences, University of Porto, Portugal, in 2007, and the M.Sc. degree in mathematical engineering from the School of Sciences, University of Porto, Portugal, in 2008. She is currently pursuing the Ph.D. degree with the School of Engineering, University of Porto, Portugal. Since 2007, she has been a Researcher at INESC TEC, an R&D Institute affiliated to University of Porto, Visual Computing and Machine Intelligence Group (VCMI). She was a Project Member of one FCT (Foundation of Science and Technology - Portugal) Research Project in the area of optical music recognition. She is currently an Assistant Professor with the Universidade Portucalense Infante D. Henrique. She is also a Senior Researcher at INESC TEC. Her main research interests include computer vision, image processing, biometrics, and document analysis.