Fusing transformed deep and shallow features (FTDS) for image-based facial expression recognition

Fares Bougourzi, Fadi Dornaika, K. Mokrani, Abdelmalik Taleb-Ahmed, Yassine Ruichek

Expert Systems With Applications 156 (2020) 113459

Fusing Transformed Deep and Shallow features (FTDS) for image-based facial expression recognition

F. Bougourzi [a, *], F. Dornaika [b, c], K. Mokrani [a], A. Taleb-Ahmed [d], Y. Ruichek [e]

[a] LTII laboratory, University of Bejaia, Algeria
[b] University of the Basque Country UPV/EHU, San Sebastian, Spain
[c] IKERBASQUE, Basque Foundation for Science, Bilbao, Spain
[d] IEMN UMR CNRS 8520 UPHF Laboratory, OAE Departement, Valenciennes, France
[e] CIAD, Univ. Bourgogne Franche-Comte, UTBM, Belfort F-90010, France

A b s t r a c t

In this paper, we propose combining between the transformed hand-crafted and deep features using PCA to recognize the six-basic facial expressions from static images. To evaluate our approach, we use three popular databases (CK+, CASIA and MMI). We introduce the use of the Pyramid Multi Level (PML) face representation for facial expression recognition. The hand-crafted features are obtained with such repre- sentations. Initially, we determine the optimal level of the PML features of three hand-crafted descriptors (HOG, LPQ and BSIF) using CK+, CASIA and MMI databases.

After the optimal level of the PML is found for each descriptor, we combine them together with the transformed final VGG-face layers (FC6 and FC7) in order to get a compact image descriptor. In within- database experiments, our approach achieved higher accuracy than the state-of-art methods on both the CK+ and CASIA databases, and competitive result on the MMI database. Likewise, our approach outper- formed the static methods in all six experiments of cross-databases.

Keywords:
Facial expression; PML; Representation; Deep features; Hand-crafted features; Cross-databases

1. Introduction

In recent years, facial expression recognition (FER) field has reached some maturity due to the considerable data augmentation and abundant methods that have reached high performance. In fact, most of the recent studies have concentrated on the recognition of the facial expression for subjects that were not included in the training phase which known as Subject-Independent protocol. Furthermore, there are some works that have studied the generalization ability of their method on cross-databases task such as Xie, Jia, Shen, and Yang (2019), and Shan, Gong, and McOwan (2009).

The facial expression recognition methods can be divided into hand-crafted and Deep Learning methods. The hand-crafted methods have been one of the most popular methods for FER. Therefore, Boosted Local Binary Pattern (Boosted-LBP) (Shan et al., 2009), Local Phase Quantization (LPQ) (Wang & Ying, 2012), Histogram of Oriented Gradients (HOG) (Carcagn, Del Coco, Leo, & Distante, 2015), Local Mean Binary Pattern (LMBP) (Goyani & Patel, 2017), and Local Directional Texture Pattern (LDTP) (Ryu, Rivera, Kim, & Chae, 2017) are some successful hand-crafted methods that have been used for facial expression recognition.

Since the publication of the first Deep Learning architecture "alexnet" (Krizhevsky, Sutskever, & Hinton, 2012), CNNs become a popular approach in FER field (Cai et al., 2019). In Lopes, de Aguiar, De Souza, and Oliveira-Santos (2017), Lopes et al. proposed a network consisting of two convolutional layers, two max pooling layers and one fully connected layer. In the training phase, the original data were augmented by using rotation to different angles and adding 2D gaussian noise in the locations of the eyes. Cai et al. (2018) proposed to use novel island loss (IL) on two CNN architectures to enhance the discriminative power of Deep Learned features. IL-CNN includes three convolutional layers. Each of the two first convolutional layers is followed by max pooling layer and the third convolutional layer is followed by FC layer then Island Loss (IL) layer. Finally, a softmax loss is used as decision layer. For the second CNN, which is denoted as IL-VGG, they fine-tuned the pre-trained VGG-Face (Parkhi, Vedaldi, & Zisserman, 2015) model with the island loss. For both architectures, data augmentation strategy was used. In Li et al. (2019), the authors proposed a network with three blocks, each block contains

* Corresponding author.
E-mail addresses: faresbougourzi@gmail.com (F. Bougourzi), fadi.dornaika@ehu.es (F. Dornaika), taleb@uphf.fr (A. Taleb-Ahmed), yassine.ruichek@utbm.fr (Y. Ruichek).

DOI : 10.1016/j.eswa.2020.113459

two convolutional layers followed by max pooling layer. The last block is followed by FC and softmax layers. To learn better feature representation and reduce the interference of noise information, they used spatial pyramid pooling model between the last convolutional layer and the second convolutional layer at different scales. Xie et al. (2019) used modified networks of VGG (Simonyan & Zisserman, 2014) and ResNet (He, Zhang, Ren, & Sun, 2016) by integrating sparse regularization into the loss function. The proposed networks with sparseness regularization achieved competitive performances in Subject-Independent protocol and proved its generalization capability in cross-databases experiments.

Although Deep Learning architectures have outperformed the facial descriptors, the CNN architectures require high computational and experimental cost. In fact, CNNs have many hyper-parameters to be optimized which makes finding these hyper-parameters a tedious task. In addition to computation cost, huge labelled data is required to train the CNNs. Indeed, the available data in FER field are limited. To solve this issue, the state-of-art methods have used data augmentation technique with various operations: translations, rotations and skewing. This makes finding the appropriate technique highly increasing the overall computation and time cost of the CNNs.

In this paper, we propose to combine the shallow and deep features to recognize the six-basic facial expressions from static images. We implemented the PML representation for the shallow descriptors to gain more sophisticated features. Moreover, our approach is strengthened by exploiting Deep Learning features. We used pre-trained VGG-face model as feature descriptor to avoid the computation complexity or the need of data augmentation, which decrease the interest of Deep Learning methods. Our proposed model exploits diversity that can be found in the shallow and deep features.

In summary, the contributions of this paper are:

- We use the ℓ-PML representation of the shallow features for FER.
- We transform both the optimal ℓ-PML features of each descriptor (HOG, LPQ and BSIF (Binarized Statistical Image Features)) and the deep features to their corresponding eigenvectors.
- We combine the transformed ℓ-PML features (HOG, LPQ, and BSIF), and the transformed deep features (FC6, FC7) by concatenating them alongside each other.

This paper is organized as follows: we illustrate our approach and the used methods in Section 2. Section 3 describes the used databases. Section 4 presents the experimental results and comparison with the state-of-art methods. Finally, we conclude our work in Section 5.

2. Methodology

2.1. Our approach

Our approach is an extension of our previous work (Bougourzi et al., 2019). In Bougourzi et al. (2019), we proposed to fuse different shallow features using PCA. After the face is aligned and cropped, we extracted three types of features which are HOG, LPQ, and BSIF. By using multi-blocks representation, we got feature histograms for each descriptor type. We then transformed each descriptor features to their corresponding eigenvectors. Finally, all transformed (compact) features are concatenated to obtain the final feature vector.

Our proposed approach is summarized in Fig. 1. The illustrated steps correspond to the test phase. During the training phase, all depicted steps are used. However, the individual PCA transformed features and the SVM model are learned using a set of labeled images. Compared to our previous work, the proposed approach presents four main differences:

1. We improved our face detection method by using more facial landmarks to assign the boundaries of the facial box.
2. We proposed to use ℓ-PML features representation of the shallow features for FER.
3. We proposed to transform the optimal ℓ-PML features of each descriptor and the deep features to their corresponding eigenvectors.
4. We combined the transformed ℓ-PML features (HOG, LPQ, and BSIF), and the transformed deep features by concatenating them alongside each other. Thus, we fused shallow and deep features in a reduced space.

2.2. Face region of interest

Since our approach is mainly based on facial texture descriptors, the face region retrieval from the raw face image, should satisfy two requirements. First, the faces should be cropped correctly to avoid missing important facial parts or adding non-facial ones. Second, the faces should be registered.[1] That is to say the facial parts have to be matched from one face to another. The process is illustrated in Fig. 2. First, we detect 68 facial landmarks using the Dlib library (King, 2009), then we exploit the 2D locations of the two eyes in order to compensate the possible in-plane rotation. After performing this 2D rotation on the image and on the detected points, the three furthest points in the left, right, and bottom direction are selected as the three boundaries of the face. We denote the distance from the lower boundary to the eyes position as d_1. The upper boundary of the face is set at a distance d_2 from the eyes that is set to $d_2 = 0.6\ d_1$. Finally, the face ROI is obtained by cropping the face using the four boundaries and re-sizing the obtained box image to a fixed size of 240×240 pixels. As consequence, the vertical position for the eyes is fixed and both vertical distances d_1 and d_2 are scaled to a constant distances D_1 and D_2, respectively.

In order to show the difference between the proposed face cropping and the classic cropping way (normalizing the intra-ocular distance that is adopted by many published works), Fig. 3 illustrates two faces that are cropped by two different schemes: the proposed one, and the classic one. From Fig. 3, we notice that the proposed scheme is more efficient than the classic one, where the cropped face contains almost all face texture and exclude non face regions as much as possible.

2.3. Descriptors

2.3.1. Histogram of oriented gradients (HOG)

HOG (Dalal & Triggs, 2005) descriptor was originally developed for human detection. Furthermore, it has been successfully used for FER (Carcagn et al., 2015). In HOG method, the input image is divided into small blocks called cells, then the occurrence of gradient orientations is counted for each cell. The cells are grouped into overlapping blocks. The concatenation of the block histograms produces the final HOG histogram.

2.3.2. Local phase quantization (LPQ)

LPQ (Ojansivu & Heikkil, 2008) is a local descriptor that uses short-term Fourier transform on local $M \times M$ neighborhoods to quantize the phase of Fourier transform by considering four frequencies. In our experiments, we choose LPQ parameters as follows: the size of the local window is 13×13 pixels, the frequency estimation method is the Gaussian derivative quadrature filter pair.

[1] This requirement is very important for hand-crafted features.

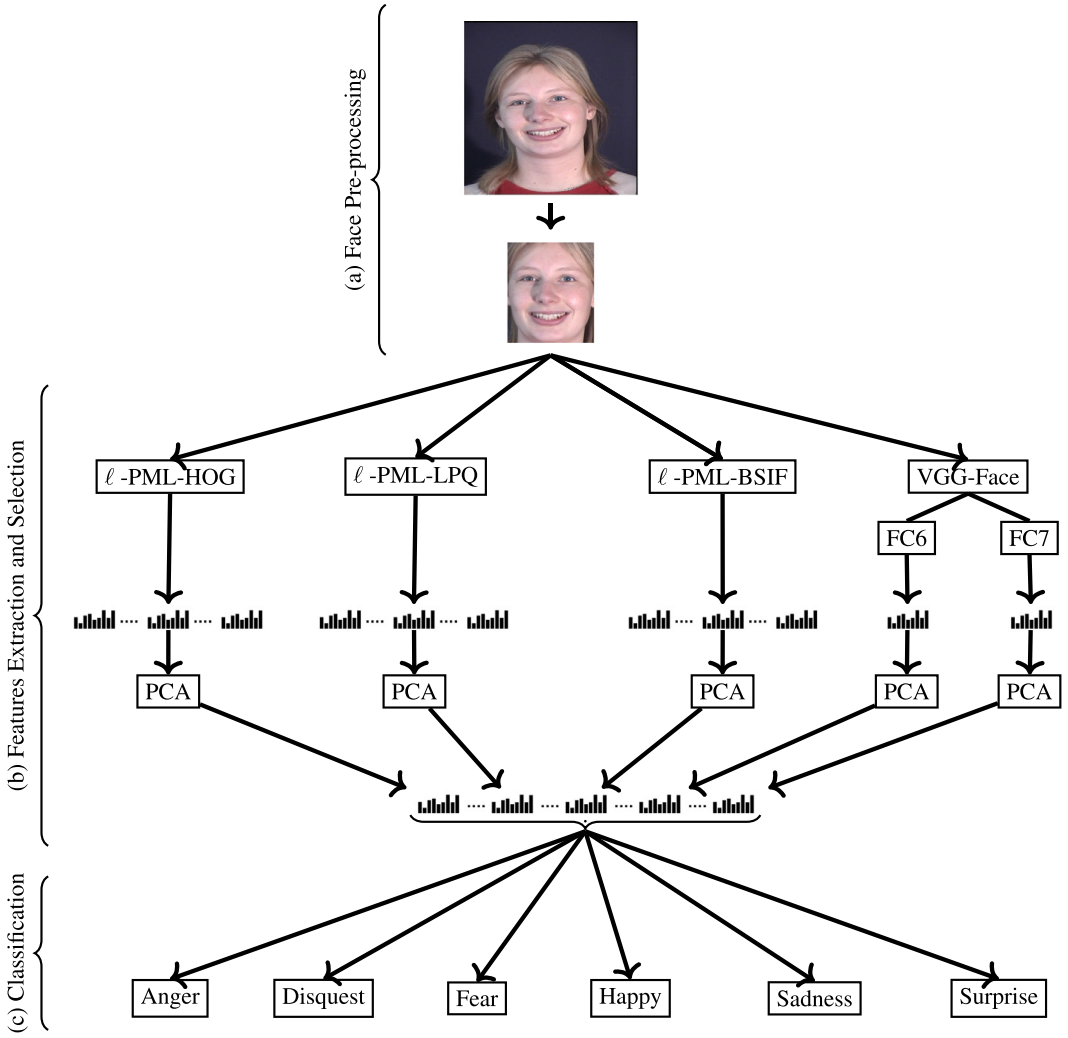

Fig. 1. General structure of the FTDS proposed approach.

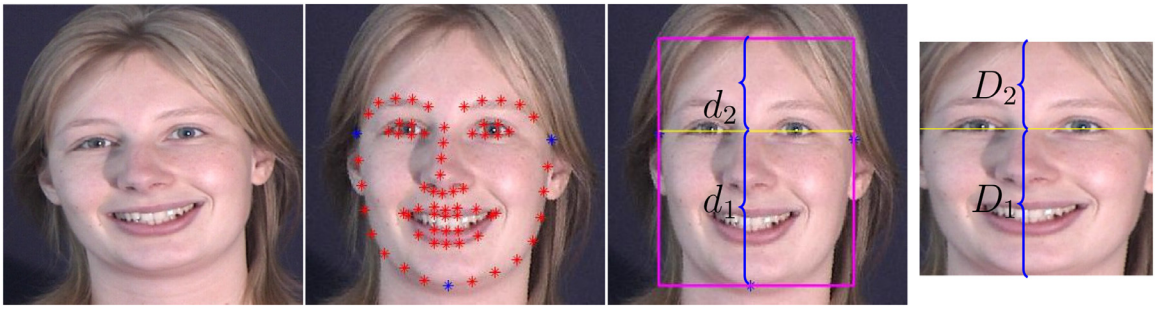

Fig. 2. Face ROI. The left image is an original image that depicts a happy expression from the MMI database. The second image is the rotated face with its detected 68 landmarks that are used to determine the three face boundaries (right, left and bottom). These boundaries correspond to the three points marked in blue *. The third image illustrates how the upper face boundary is obtained. It is placed at a distance $d_2 = 0.6 \times d_1$ from the vertical position of the eyes. The fourth image illustrates the cropped and re-sized face image of 240 × 240 pixels. Note that the distances D_1 and D_2 are constant for all cropped faces. (For a better illustration of colours, the reader is referred to the web version of this article.)

2.3.3. Binarized statistical image features (BSIF)

BSIF (Kannala & Rahtu, 2012) features are obtained by binarizing the response of filtering the input image by a set of predefined 2D filters. The filters are learned from natural images using independent component analysis. In our experiments, we chose 17 × 17 × 8 bank of filters.

2.4. PML Representation

In order to extract rich and sophisticated texture description, our approach is based on the PML (Bekhouche, Ouafi, Dornaika, Taleb-Ahmed, & Hadid, 2017). The idea of ℓ-PML is to generate a physical pyramid having ℓ levels and divide these ℓ images into different blocks appropriately. In fact, ℓ-PML representation contains

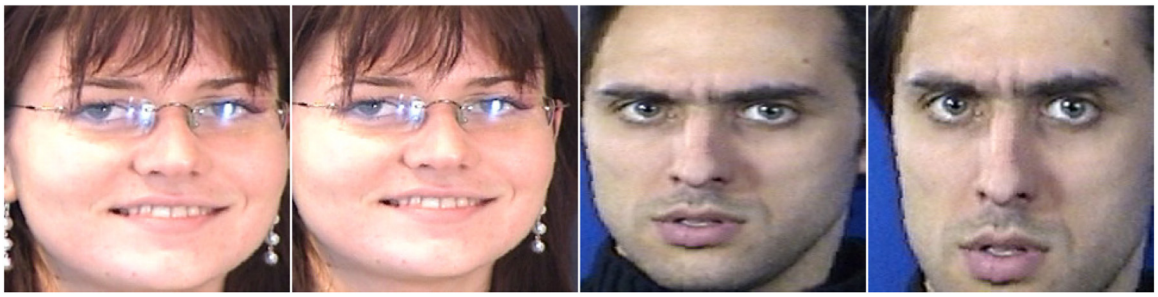

Fig. 3. Face detection comparison. The proposed face detection examples are shown in the first and the third images. While, the second and the fourth images are the corresponding detected faces using the classic scheme.

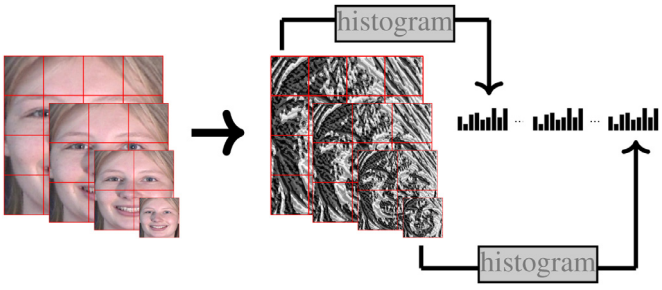

Fig. 4. Pyramid Multi-Level Binarized Statistical Image Features example of 4 levels.

ℓ levels. We denote the ith level by L_i. Thus, the pyramid consists of $\{L_1, L_2, ..., L_\ell\}$. For a squared image of size $n \times n$, the basic block size of ℓ-PML is $s = n/\ell$. The size of L_i is $[i \times s, i \times s]$ and its block division is $i \times i$. By following the previous relation, the size of L_1 (coarsest image) is $[s, s]$ and its block division is 1×1, and the size of L_ℓ (the finest image) is $[n, n]$ and its block division is $\ell \times \ell$.

The final descriptor of ℓ-PML is formed by the concatenation of the descriptors of all blocks in all levels. The number of the concatenated descriptors is equal to the total number of the blocks $N_b = \sum_{i=1}^{\ell} i^2 = \frac{\ell(\ell+1)(2\ell+1)}{6}$. Fig. 4 shows an example of 4-PML-BSIF.

2.5. Deep features

For deep face features, we used the pre-trained deep CNN model VGG-Face (Parkhi et al., 2015). This model was originally developed for face recognition tasks. It was trained on over 2.6K persons. VGG-Face has the ability to extract different facial features, since it was trained on huge number of facial images. We selected the FC6 and FC7 layers as two feature descriptors. We obtained a feature vector of size 4096 from each layer.

2.6. Features transformation using PCA

In the literature, PCA (Martnez & Kak, 2001) is used to cope with the curse of dimensionality and to extract the most discriminative features from the raw ones. For an n-by-d data matrix X (n is the number of samples and d is the number of features), PCA computes an effective linear transformation by taking the eigenvectors of the covariance matrix $Q = X^T X$. This is obtained by solving the eigen equation defined by: $Q\, e_i = \lambda_i\, e_i$, where λ_i is the eigenvalue corresponding to the eigenvector e_i. In our work, we consider the eigenvectors of XX^T as the transformation of our raw data matrix X. The eigenvectors are ranked by the order of their eigenvalues, from highest to lowest eigenvalue. In our experiments, we grid search about the optimal number of transformed features.

2.7. Facial expression recognition using SVM

After we found the Final-PCA-Fusion descriptor (concatenation of the ℓ-PML and deep features using the PCA outputs),we used lib-linear library (Fan, Chang, Hsieh, Wang, & Lin, 2008) based linear-SVM classifier in order to learn our classifier without any hyper-parameter tuning.

3. Databases

To evaluate the performance of our approach, we used three public databases: CK+, CASIA and MMI:

CK+ (Lucey et al., 2010) is a facial expression database that consists of 593 image-sequences from 123 subjects. The database is designed for the six-basic expressions (Anger, Disgust, Fear, Happy, Sadness and Surprise) plus Contempt. In our experiments, we selected the three last frames (peak frames) for the six-basic expressions, therefore we collected 927 images from 106 persons. The images are in gray-scale or color and their resolution is 640×490 or 640×480 pixels.

Oulu-CASIA (Zhao, Huang, Taini, Li, & PietikInen, 2011) NIR VIS database consists of 80 persons where each person has six image-sequences corresponding to the six-basic expressions. Each sequence begins with the neutral face and ends with the peak of the corresponding expression. We selected the last three frames from each sequence as we did with CK+ database. In total, we got 1440 color images with resolution of 320×240 pixels.

MMI (Valstar & Pantic, 2010) consists of 208 videos that are labeled as one of the six-basic expressions. These sequences are obtained from 31 persons. In contrast to CK+ and CASIA, MMI peak expression frames are unknown. MMI's clips begin with the neutral face and end with it and the expression is in between. We approximated three peak frames in the middle of each clip, so we obtained 624 color images with resolution of 720×576 pixels.

In the collection of the static images from image sequences, we followed the research community protocol, in which the last three frames of each sequence are selected for CK+ and CASIA Cai et al. (2018); Kuo, Lai, and Sarkis (2018); Xie et al. (2019) and the three middle frames are selected for MMI database (Cai et al., 2019; 2018; Ryu et al., 2017).

4. Performance evaluation

To evaluate the performance of our approach, we conducted two types of experiments: within-database classification and cross-databases classification. For within-databases experiments, we used Leave-One-Subject-Out (LOSO) scheme. The LOSO scheme is a K-folds Subject-Independent Cross-Validation (K-folds SI CV) protocol where the number of folds (K) is equal to the number of subjects; for each fold we consider one subject's samples as the testing data and the rest subject's samples will be used as training

Table 1
Classification accuracy (%) as a function of the number of PML levels for each descriptor on the CK+, CASIA and MMI databases using LOSO protocol. N* is the number of selected PCA features corresponding to the order of their eigenvalues, from highest to lowest eigenvalue.

Descriptor	ℓ	CK+	N*	CASIA	N*	MMI	N*
	12	**97.20**	**100**	**78.40**	**140**	**68.44**	**30**
	10	96.55	130	78.01	170	65.83	40
PML-HOG	8	96.22	130	76.49	190	67.07	40
	6	94.39	130	74.96	190	67.71	60
	4	93.31	150	68.08	180	63.94	50
	12	96.12	150	83.24	180	65.38	40
	10	96.55	150	83.14	160	65.71	50
PML-LPQ	**8**	**97.20**	**120**	**83.89**	**190**	**67.47**	**120**
	6	97.09	130	83.25	170	66.02	50
	4	95.90	120	82.33	150	67.43	50
	12	96.66	140	83.61	210	60.90	50
	10	96.66	140	84.31	200	61.06	50
PML-BSIF	8	97.41	150	84.72	210	62.66	150
	6	**97.52**	**150**	**85.14**	**200**	**63.45**	**60**
	4	93.85	150	82.15	240	59.46	50

data. For the cross-databases experiments, all the three databases are considered. Each time, one database is used as testing data and another one is used as training data. This cross-databases experiments seek to quantify the generalization ability of our approach.

4.1. Within database experiment

Our proposed approach relies on the use of ℓ-PML, which has the number of levels as a parameter. In order to determine this parameter and fix it for subsequent use, we used CK+, CASIA and MMI databases to find the optimal number of levels for each hand-crafted descriptor. To this end, we apply our scheme on the individual descriptors HOG, LPQ and BSIF. We used 4, 6, 8, 10 and 12 PML levels. Table 1 depicts the obtained results. From Table 1, we found that the optimal number of PML-levels are 12, 8 and 6 for HOG, LPQ and BSIF, respectively, for all the evaluated databases. After we found the optimal number of PML-levels for each descriptor, we evaluated transformed individual features (three hand-crafted types and two deep types) and four fusion schemes on CK+, CASIA and MMI databases. For the fusion schemes, we concatenated equal number of transformed individual features. The optimal number of PCA features is obtained by grid search strategy. The results are summarized in Table 2.

From Table 2, we can notice that the fusion between the optimal ℓ-PML (HOG, LPQ, and BSIF) and deep features achieved the highest accuracy compared with the scheme that uses an optimal ℓ-PML of one single descriptor on CK+, CASIA, and MMI databases. Although the deep features (both FC6 and FC7 of VGG-Face model) gave lower accuracy than the ℓ-PML of a single descriptor, fusing them with the ℓ-PML improves the performance. This occurred because the features type of deep features is different of the hand-crafted ones. The proposed approach (FTDS) gave better performance than the best individual feature by 1.07%, 4.51% and 5.63% for CK+, CASIA and MMI databases, respectively. This proves the efficiency of combining different feature types (three hand-crafted types and two deep types) for recognizing the facial expressions in different databases.

Table 3 depicts all the specifics of our approach and those related to the state-of-art works we considered for comparison on the evaluated databases (CK+, CASIA and MMI). The specifics includes: the type of approach (Deep Learning, Shallow or Deep + Shallow) and the detailed information of the used databases (number of subjects, number of selected frames from a sequence and number of expressions).

Tables 4, 5 and 6 illustrate the comparison between our proposed approach and several recent state-of-art methods on CK+, CASIA and MMI, respectively. All of the comparisons to the state-of-art methods were conducted using Subject-Independent (SI) protocol. As we can see from the fourth column of the Tables 4, 5 and 6, there is no unique number of K-folds SI CV in the research community. Due to that, we conducted our experiments using LOSO scheme (K equals the number of subjects of the evaluation database) to avoid random grouping of the subjects into folds which will be changed from one experiment to another as explained in our previous paper (Bougourzi et al., 2019).

Our approach gives better performance than our previous work by 2.31%, 9.66% and 0.5% for the CK+, CASIA and MMI databases as shown in the Tables 4, 5 and 6, respectively. Furthermore, our approach achieved higher accuracy than the state-of-art methods on the CK+ and CASIA databases as shown in Tables 4, and 5, respectively. For CK+ database, we included, in the comparison table (Table 4), two methods that have used fusion between shallow and deep methods (Sun & Lv, 2019; Zeng, Zhou, Jia, Xie, & Shen, 2018), despite that they have not used exactly the same subjects images as in other published works (has not selected the same number of frames from each sequence) as depicted in Table 3.

On the other hand, our proposed approach is among the best approaches on MMI database as illustrated in Table 6, where the best method outperformed ours by only 0.6%. Similar to CK+ database, we included, in the comparison table (Table 6), a method that has used fusion between shallow and deep methods (Zeng et al., 2018), despite it has not used exactly the same subjects images as in other published works (has not selected the same number of frames from each sequence and did not use all the available subjects) as depicted in Table 3.

Table 2
Performance evaluation using transformed individual features (three hand-crafted types and two deep types) and fused features on CK+, CASIA and MMI databases using LOSO protocol. Four fusion schemes are tested. ψ denotes the fusion between PML-HOG, PML-LPQ, PML-BSIF, VGG-FC7 and VGG-FC6 using PCA. N* is the optimal number of PCA features. In the experiments of features fusion, N* is the optimal number of PCA features from each descriptor.

Features Type \ Database	CK+		CASIA		MMI	
	Accuracy	N*	Accuracy	N*	Accuracy	N*
12-PML-HOG (1)	97.20	100	78.40	140	68.44	30
8-PML-LPQ (2)	97.20	120	83.89	190	67.47	50
6-PML-BSIF (3)	97.52	150	85.14	200	63.45	60
VGG-FC7 (4)	90.83	90	75.63	160	62.96	60
VGG-FC6 (5)	94.39	150	80	210	66.34	50
(1) + (2) + (3)	**98.27**	**150**	87.15	230	70.37	50
(1) + (2) + (3) + (4)	97.30	150	86.18	230	72.62	50
(1) + (2) + (3) + (5)	97.63	130	87.92	210	72.46	50
FTDS ψ (Prop. approach)	**98.27**	**120**	**89.65**	**190**	**74.07**	**50**

Table 3
The specifics of our approach and those related to the state-of-art works we considered for comparison on the evaluated databases (CK+, CASIA and MMI). * is the number of subjects; ** is the number of selected frames from a sequence; *** is the number of expressions.

Database	Article	Method Type	N. Sub.*	N. F/S**	N. Exp. ***
	Xie et al. (2019)	Deep Learning	106	3 last frames	6
	Li et al. (2019)	Deep Learning	106	3 last frames	6
	Yang, Ciftci, and Yin (2018a)	Deep Learning	106	3 last frames	6
	Lopes et al. (2017)	Deep Learning	106	3 last frames	6
	Cai et al. (2019)	Deep Learning	106	3 last frames	6
	Yang, Zhang, and Yin (2018b)	Deep Learning	106	3 last frames	6
CK+	Cai et al. (2018)	Deep Learning	106	3 last frames	6
	Ryu et al. (2017)	Shallow	106	3 last frames	6
	Cai et al. (2018)	Deep Learning	106	3 last frames	6
	Zeng et al. (2018)	Shallow + Deep	106	1 last frame	6
	Sun and Lv (2019)	Shallow + Deep	106	first + 3 last frames	7
	Zhang et al. (2019)	Deep Learning	106	video-based (6 frames)	6
	Ours	Shallow+Deep	106	3 last frames	6
	Yang et al. (2018b)	Deep Learning	80	3 last frames	6
	Kuo et al. (2018)	Deep Learning	80	3 last frames	6
	Yang et al. (2018a)	Deep Learning	80	3 last frames	6
CASIA	Otberdout, Kacem, Daoudi, Ballihi, and Berretti (2018)	Deep Learning	80	3 last frames	6
	Cai et al. (2018)	Deep Learning	80	3 last frames	6
	Xie et al. (2019)	Deep Learning	80	3 last frames	6
	Zhang et al. (2019)	Deep Learning	80	video-based (6 frames)	6
	Ours	Shallow+Deep	80	3 last frames	6
	Li et al. (2019)	Deep Learning	20	2 middle frames	6
	Cai et al. (2018)	Deep Learning	31	3 middle frames	6
	Cai et al. (2019)	Deep Learning	31	3 middle frames	6
MMI	Yang et al. (2018a)	Deep Learning	31	3 middle frames	6
	Ryu et al. (2017)	Shallow	28	3 middle frames	6
	Sun and Lv (2019)	Shallow + Deep	20	1 middle frame	6
	Zhang et al. (2019)	Deep Learning	30	video-based (6 frames)	6
	Ours	Shallow+Deep	31	3 middle frames	6

Table 4
Comparison with the state-of-art methods on CK + database using Subject-Independent protocol. * Our previous method; ** the work reported in that paper has not used exactly the same subjects images as in other published works.

Article	Method	Accuracy	Protocol
Xie et al. (2019)	Sparse CNN	97.59	10-folds SI CV
Li et al. (2019)	C-SPP	97.41	8-folds SI CV
Kuo et al. (2018)	Compact CNN	97.37	10-folds SI CV
Yang et al. (2018a)	DeRL	97.30	10-folds SI CV
Lopes et al. (2017)	CNN	96.76	8-folds SI CV
Yang et al. (2018b)	CNN	96.57	LOSO
Cai et al. (2019)	IF-GAN	95.90	10-folds SI CV
Cai et al. (2018)	IL-CNN	94.39	LOSO
Ryu et al. (2017)	LDTP	94.2	LOSO
Cai et al. (2018)	IL-VGG	91.64	LOSO
Zeng et al. (2018)**	GSF+CNN	97.35	10-folds SI CV
Sun and Lv (2019) **	CNN-SIFT + SVM	94.13	8-folds SI CV
Bougourzi et al. (2019) (Pr)*	MB-fusion-PCA	95.96	LOSO
Proposed approach	**FTDS**	**98.27**	LOSO

Table 5
Comparison with the state-of-art methods on CASIA database using Subject-Independent protocol. * Our previous method.

Article	Method	Accuracy	Protocol
Yang et al. (2018b)	CNN	88.92	LOSO
Kuo et al. (2018)	Compact CNN	88.75	10-folds SI CV
Yang et al. (2018a)	DeRL	88.00	10-folds SI CV
Otberdout et al. (2018)	ExpNet	84.80	LOSO
Cai et al. (2018)	IL-VGG	84.58	LOSO
Xie et al. (2019)	Sparse CNN	82.71	10-folds SI CV
Cai et al. (2018)	IL-CNN	77.29	LOSO
Bougourzi et al. (2019) (Pr)*	MB-fusion-PCA	79.99	LOSO
Proposed approach	**FTDS**	**89.65**	LOSO

4.2. Cross databases experiment

In addition to intra-database experiments, we evaluated the performance of our approach in cross-databases tasks. In total, we produced 6 cross-databases experiments, in each one we selected one database as a training database and another one as the testing one. There are many factors that control the cross-databases results such as the caption conditions of each database images which include illumination, resolution and occlusion. The number of subjects and the balance of the training database also have great effect. Table 7 summarizes the results of the six cross-databases experiments using the individual optimal PML features (i.e., 12-PML-HOG, 8-PML-LPQ and 6-PML-BSIF). From the results, we observe that there is no individual feature type that performs better than the others in all cross-databases experiments. In more details, we

Table 6
Comparison with the state-of-art methods on MMI database using Subject-Independent protocol. * Our previous method; ** the work reported in that paper has not used exactly the same subjects images as in other published works.

Article	Method	Accuracy	Protocol
Cai et al. (2018)	IL-VGG	74.68	LOSO
Cai et al. (2019)	IF-GAN	74.52	10-folds SI CV
Yang et al. (2018a)	DeRL	73.23	10-folds SI CV
Cai et al. (2018)	IL-CNN	70.67	LOSO
Ryu et al. (2017)	LDTP	67.86	LOSO
Li et al. (2019) **	C-SPP	59.20	7-folds SI CV
Sun and Lv (2019) **	CNN-SIFT + SVM	53.81	8-folds SI CV
Bougourzi et al. (2019) (Pr)*	MB-fusion-PCA	73.57	LOSO
Proposed approach	**FTDS**	**74.07**	LOSO

Table 7
Cross-databases experiments using individual transformed features (three hand-crafted types). N* is the number of selected features corresponding to the order of their eigenvalues, from highest to lowest eigenvalue.

Database\Features Type		12-PML-HOG		8-PML-LPQ		6-PML-BSIF	
Tr. data	Ts. data	Accuracy	N*	Accuracy	N*	Accuracy	N*
CASIA	CK+	72.32	30	78.32	40	87.92	30
MMI	CK+	78.43	30	67.32	30	64.08	30
CK+	MMI	63.62	30	56.41	50	52.08	40
CASIA	MMI	59.62	40	60.58	70	56.09	80
CK+	CASIA	50.21	50	59.58	40	63.61	60
MMI	CASIA	55.14	30	56.11	40	51.39	40

observe that 12-PML-HOG gave the best accuracy when CK+ and MMI are used as the training and testing databases and as testing and training databases. Also, 8-PML-LPQ gave the best accuracy when CASIA and MMI are used as training and testing databases and as testing and training databases. When CASIA and CK+ are used as training and testing databases and as testing and training databases, 6-PML-BSIF gave the best accuracy.

Table 8 summarizes the results of the six cross-databases experiments using the two deep features (FC6 and FC7) and our FTDS approach. The comparison with the results in Table 7 shows that the deep features achieved competitive performance with the individual descriptors. Also, the fusion between all features (FTDS) gave the best performance in all of the six experiments of cross-databases, that proves the efficiency of our method on combining different features types, taking the advantage from each type and reducing their drawbacks. Also, we can observe that our approach (FTDS) can achieve high accuracy with lower feature dimension than the raw features.

From the FTDS results (fifth column of Table 8), the performances when CK+ is the testing database are higher than those obtained when MMI or CASIA is the testing one. This occurred because CK+ is less challenging than MMI and CASIA. From the first and second rows of the fifth column of Table 8, we observe that using CASIA as training database gives higher accuracy than when using MMI as training database. This happened because CASIA database contains higher number of subjects and it is more balanced than MMI database. From the third and fourth rows of the fifth column of Table 8, we observe that using CASIA as training database gives less accuracy than when using CK+. This is due to the considerable difference in image resolution between the training (CASIA) and testing (MMI) databases. The image resolution in CASIA database is 320 $\times$ 240 pixels and that of MMI database is 720 $\times$ 576 pixels. The same phenomenon occurs when MMI is used as training database and CASIA as the testing one, as shown in the fifth and sixth rows of the fifth column of Table 8.

Since there are few works that have tested their methods on Cross-databases tasks, we compared our approach with the state-of-art methods that were trained and tested on the same databases as we did. The comparison results are summarized in Table 9 which includes two methods Sun and Lv (2019) and Zhang, Mahoor, and Mavadati (2015) that produced just one and two cross-databases experiments, respectively. From this table, we observe that our method achieved higher performance in these experiments. Although the work described in Sun and Lv (2019) has not used exactly the same subjects images as in other published works (has not selected the same number of frames from each sequence and did not use all the available subjects of MMI database) as depicted in Table 3, we made the comparison with it because they used fusion between shallow and deep learning methods as we did.

In addition, Table 9 includes one static method (Xie et al., 2019), and an extra video-based method (Zhang, Xia, & Liu, 2019) which made the six cross-databases experiments as we did. The video-based methods use sequence of images as input instead of one single image and classify them entirely. To the best of our knowledge, Xie et al. (2019) and Zhang et al. (2019) are the only available works that tested their method on the six cross-databases experiments as we did. From the comparison of these results, we observe that our approach outperforms the static method in all six exper-

Table 8
Cross-databases experiments using individual transformed features (two deep types) and the proposed fused features (The three hand-crafted types and the two deep types). N* is the optimal number of PCA features. In the experiments of features fusion, N* is the optimal number of PCA features from each descriptor.

Database\Features Type		VGG-FC6		VGG-FC7		FTDS	
Tr. data	Ts. data	Accuracy	N*	Accuracy	N*	Accuracy	N*
CASIA	CK+	74.97	30	72.06	50	**88.35**	**20**
MMI	CK+	63.86	30	58.14	30	**78.96**	**20**
CK+	MMI	61.22	80	61.22	100	**67.63**	**100**
CASIA	MMI	59.14	70	57.53	90	**62.34**	**100**
CK+	CASIA	59.60	50	56.32	70	**68.33**	**70**
MMI	CASIA	53.89	50	48.55	70	**63.06**	**50**

Table 9
Comparison of the proposed approach with the state-of-art methods on Cross-Databases. * the work reported in that paper has not used exactly the same subjects images as in other published works; ** video-based framework.

Tr. data	Ts. data	Sun and Lv (2019)*	Zhang et al. (2015)	Xie et al. (2019)	Zhang et al. (2019)**	Ours
CASIA	CK+	-	-	84.47	64.72	**88.35**
MMI	CK+	-	61.2	77.02	58.90	**78.96**
CK+	MMI	53.81	66.9	60.48	**67.80**	67.63
CASIA	MMI	-	-	61.46	60.49	**62.34**
CK+	CASIA	-	-	42.08	61.46	**68.33**
MMI	CASIA	-	-	50.83	**76.10**	63.06

iments. Also, our approach performs better than the video-based method in four experiments out of six.

5. Conclusion

In this paper, we presented a framework for fusing the hand-crafted features with deep features. The resulting framework can compete with pure Deep Learning architectures. The used hand-crafted features adopted sophisticated face representations such as PML. Moreover, the fusion process is performed in transformed subspaces such as PCA. Our approach is not only designed to combine different hand-crafted methods, but it allows to combine the shallow and deep features. The proposed approach performed better than most of the State-of-art methods on both the within-database and cross-databases experiments. As future work, we propose to exploit more descriptors and pre-trained CNN architectures. In addition, we will extend our work on evaluating a committee of classifiers. We also plan to fit our approach to recognize the facial expressions from videos.

Declaration of Competing Interest

We wish to confirm that there are no known conflicts of interest associated with this publication and there has been no significant financial support for this work that could have influenced its outcome.

Credit authorship contribution statement

F. Bougourzi: Software, Validation, Resources, Investigation, Data curation, Writing - original draft, Writing - review & editing, Visualization. **F. Dornaika:** Conceptualization, Formal analysis, Investigation, Resources, Writing - original draft, Writing - review & editing, Supervision, Project administration. **K. Mokrani:** Resources, Supervision. **A. Taleb-Ahmed:** Resources, Supervision. **Y. Ruichek:** Conceptualization, Supervision, Writing - review & editing.

Acknowledgments

This work was funded by the Spanish Ministerio de Ciencia, Innovacion y Universidades, Programa Estatal de I+D+i Orientada a los Retos de la Sociedad, RTI2018-101045-B-C21.

References

Bekhouche, S. E., Ouafi, A., Dornaika, F., Taleb-Ahmed, A., & Hadid, A. (2017). Pyramid multi-level features for facial demographic estimation. *Expert Systems with Applications, 80*, 297–310.

Bougourzi, F., Mokrani, K., Ruichek, Y., Dornaika, F., Ouafi, A., & Taleb-Ahmed, A. (2019). Fusion of transformed shallow features for facial expression recognition. *IET Image Processing, 13*(9), 1479–1489.

Cai, J., Meng, Z., Khan, A. S., Li, Z., O'Reilly, J., & Tong, Y. (2019). Identity-free facial expression recognition using conditional generative adversarial network. arXiv preprint arXiv:1903.08051,.

Cai, J., Meng, Z., Khan, A. S., Li, Z., O'Reilly, J., & Tong, Y. (2018). Island loss for learning discriminative features in facial expression recognition. In *2018 13th IEEE international conference on automatic face gesture recognition (FG 2018)* (pp. 302–309). doi:10.1109/FG.2018.00051.

Carcagn, P., Del Coco, M., Leo, M., & Distante, C. (2015). Facial expression recognition and histograms of oriented gradients: A comprehensive study. *SpringerPlus, 4*(1), 645.

Dalal, N., & Triggs, B. (2005). Histograms of oriented gradients for human detection.

Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R., & Lin, C.-J. (2008). LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research, 9*(Aug), 1871–1874.

Goyani, M. M., & Patel, N. (2017). Recognition of facial expressions using local mean binary pattern. *ELCVIA, 16*(1), 54–67.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *2016 IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 770–778). doi:10.1109/CVPR.2016.90.

Kannala, J., & Rahtu, E. (2012). Bsif: Binarized statistical image features. In *Proceedings of the 21st international conference on pattern recognition (ICPR2012)* (pp. 1363–1366). IEEE.

King, D. E. (2009). Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research, 10*(Jul), 1755–1758.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1097–1105).

Kuo, C.-M., Lai, S.-H., & Sarkis, M. (2018). A compact deep learning model for robust facial expression recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops* (pp. 2121–2129).

Li, L., Yuan, Y., Li, M., Xu, H., Li, R., & Lu, S. (2019). Subject independent facial expression recognition: Cross-connection and spatial pyramid pooling convolutional neural network. In *Proceedings of the 2019 international conference on image, video and signal processing* (pp. 85–92). ACM.

Lopes, A. T., de Aguiar, E., De Souza, A. F., & Oliveira-Santos, T. (2017). Facial expression recognition with convolutional neural networks: Coping with few data and the training sample order. *Pattern Recognition, 61*, 610–628.

Lucey, P., Cohn, J. F., Kanade, T., Saragih, J., Ambadar, Z., & Matthews, I. (2010). The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *2010 IEEE computer society conference on computer vision and pattern recognition-workshops* (pp. 94–101). IEEE.

Martnez, A. M., & Kak, A. C. (2001). Pca versus lda. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 23*(2), 228–233.

Ojansivu, V., & Heikkil, J. (2008). Blur insensitive texture classification using local phase quantization. In *International conference on image and signal processing* (pp. 236–243). Springer.

Otberdout, N., Kacem, A., Daoudi, M., Ballihi, L., & Berretti, S. (2018). Deep covariance descriptors for facial expression recognition. arXiv:1805.03869 [cs], ArXiv: 1805.03869.

Parkhi, O. M., Vedaldi, A., & Zisserman, A. (2015). Deep face recognition.. In *bmvc: 1* (p. 6).

Ryu, B., Rivera, A. R., Kim, J., & Chae, O. (2017). Local directional ternary pattern for facial expression recognition. *IEEE Transactions on Image Processing, 26*(12), 6006–6018.

Shan, C., Gong, S., & McOwan, P. W. (2009). Facial expression recognition based on local binary patterns: A comprehensive study. *Image and Vision Computing, 27*(6), 803–816.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556 [cs], ArXiv: 1409.1556.

Sun, X., & Lv, M. (2019). Facial expression recognition based on a hybrid model combining deep and shallow features. *Cognitive Computation, 11*(4), 587–597.

Valstar, M., & Pantic, M. (2010). Induced disgust, happiness and surprise: an addition to the mmi facial expression database. In *Proc. 3rd Intern. Workshop on EMOTION (satellite of LREC): Corpora for Research on Emotion and Affect* (p. 65). Paris, France.

Wang, Z., & Ying, Z. (2012). Facial expression recognition based on local phase quantization and sparse representation. In *2012 8th international conference on natural computation* (pp. 222–225). IEEE.

Xie, W., Jia, X., Shen, L., & Yang, M. (2019). Sparse deep feature learning for facial expression recognition. *Pattern Recognition*, 106966.

Yang, H., Ciftci, U., & Yin, L. (2018a). Facial expression recognition by de-expression residue learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2168–2177).

Yang, H., Zhang, Z., & Yin, L. (2018b). Identity-adaptive facial expression recognition through expression regeneration using conditional generative adversarial networks. In *2018 13th IEEE international conference on automatic face gesture recognition (FG 2018)* (pp. 294–301). doi:10.1109/FG.2018.00050.

Zeng, G., Zhou, J., Jia, X., Xie, W., & Shen, L. (2018). Hand-crafted feature guided

deep learning for facial expression recognition. In *2018 13th IEEE international conference on automatic face & gesture recognition (fg 2018)* (pp. 423–430). IEEE.
Zhang, X., Mahoor, M. H., & Mavadati, S. M. (2015). Facial expression recognition using lp-norm MKL multiclass-SVM. *Machine Vision and Applications, 26*(4), 467–483.
Zhang, Y.-F., Xia, T., & Liu, Y. (2019). 3d convolution network and Siamese-attention mechanism for expression recognition. *Multimedia Tools and Applications*, 1–17.
Zhao, G., Huang, X., Taini, M., Li, S. Z., & PietikInen, M. (2011). Facial expression recognition from near-infrared videos. *Image and Vision Computing, 29*(9), 607–619.