New framework for person-independent facial expression recognition combining textural and shape analysis through new feature extraction approach

M. Kas, Y. El Merabet, Y. Ruichek, R. Messoussi

Version of Record: https://www.sciencedirect.com/science/article/pii/S0020025520310707
Manuscript_e97b18504f0f5d3517eff5f6c287b125

New Framework for Person-independent Facial Expression Recognition
Combining textural and shape analysis through new feature extraction approach

M. Kas[a,b,*], Y. El merabet[b], Y. Ruichek[a], R. Messoussi[b]

[a] *CIAD UMR 7533, Univ. Bourgogne Franche-Comté, UTBM, F-90010 Belfort, France*
[b] *Laboratoire LASTID, Département de Physique, Faculté des Sciences, Université Ibn Tofail, BP 133, 14000 Kenitra, Maroc*

Abstract

Automatic facial expression recognition (FER) has been extensively studied owing to its wide range of applications, such as in e-learning platforms used to automatically collect the feedback of students regarding a particular content and to help children with autism have a better understanding of their environment. Owing to the advances made in the fields of machine learning and computational devices, researchers are developing more accurate and robust facial expression recognition frameworks. In this paper, we propose a completely new framework for person-independent FER based on combining textural and shape features from 49 detected landmarks in an input facial image. The shape information is extracted using the histogram of oriented gradients (HOG) applied on a binary patch generated by interpolating the locations of the 49 detected landmarks. The textural information is computed from 49 sub-images, each centered on one landmark, using a new handcrafted descriptor that we also propose herein and is referred to as Orthogonal and Parallel-based Directions Generic Quad Map Binary Patterns (OPD-GQMBP). OPD-GQMBP encodes the relevant information based on the orthogonality and parallelism of the geometries to select the prominent pixels within a $n \times n$ neighborhood. The proposed framework outperforms many previous state-of-the-art methods including deep-learning-based approaches on five widely used benchmarks: CK+, KDEF, JAFFE, Oulu-Casia VIS, and RaFD, through the Leave-One-Subject-Out evaluation protocol. In addition, the superiority of the OPD-GQMBP descriptor is fairly proven against 10 deep features (e.g., VGG, ResNets, DenseNet, GoogeLeNet, and Inception) and 12 recent and powerful LBP variants.

Keywords: Facial expressions recognition; Textural and shape analysis; Landmarks; Local Binary Patterns; Support vector machines; Deep features;

1. Introduction

Believing that computing devices can autonomously perform complicated tasks by being trained rather than programmed, machine learning relies on endowing machines with the cognitive skills naturally acquired by the human brain. Machine learning and artificial intelligence (AI) are dominant topics in terms of how advanced analytics will appear in the future as well as their expected outcomes and benefits. More precisely, ongoing studies are oriented toward the development of machine-learning-based algorithms capable of fulfilling brain functions by allowing them to iteratively learn from data to improve, describe data, and predict outcomes. The perception of the human brain is the key to understanding and interacting with the environment based on hearing, touching, and in particular, natural vision sensors. Similarly, through state-of-the-art computer vision, researchers are attempting to implement humans-like visual analysis capabilities. One of the most complicated human visual analyses is facial expression recognition, which relies on sensing the emotion of a given person within the environment based on the individual's facial expression. This task serves diverse applications that are of interest to many different markets. For example, it can be useful to help children diagnosed with autism better understand their social environment. Moreover, such technology will allow an accurate real-time evaluation of E-learning content and public services to be more easily achieved. Furthermore, the robot industry will be able to develop human-support robots qualified to adapt their interactions according to the emotional atmosphere. In real-world scenarios, the desired system is expected to recognize the emotion of unseen individuals in real time, which makes this task among the most difficult in computer vision. In the literature, the recognition of facial expression has four different levels, as can be seen in Figure 1, which demonstrates the level of difficulty regarding the ways in which emotions are expressed (spontaneous versus posed) and the person expressing it (the same person used in the training or a different individual). Spontaneous emotions are difficult to classify because each individual expresses a given emotion differently compared to another person. Furthermore, this fact often leads to interclass sample interference, meaning that two emotion classes are represented over two images with the same overall appearance. Moreover, the recording of spontaneous emotions must be applied while the subjects are un-

*Corresponding author
Email addresses: mohamed.kas@utbm.fr (M. Kas), y.el-merabet@univ-ibntofail.ac.ma (Y. El merabet), yassine.ruichek@utbm.fr (Y. Ruichek), messoussi@uit.ac.ma (R. Messoussi)

aware of it, which is extremely difficult to establish because the subjects should deliver authorization for the recording and use of their images/videos. Therefore, there are only a few studies that have been interested in spontaneous emotion recognition and have focused solely on verifying the assigned labels and whether they match the corresponding observation, as reported by [15]. The majority of available databases for spontaneous facial expressions have been collected from the web, based on saving images from search engines (mainly Google and Flicker) by specifying the emotion-related keywords. The well-known and widely used databases of this kind are FER and AFEW. By contrast, posed expressions are obtained by requesting the subjects to perform the facial expressions in a uniform way in order to avoid intra-class similarities that would confuse the classification task. The subjects are usually skilled persons (actors), and thus their expressions can be computationally classified. The second challenge of facial expression recognition relies on correctly decoding the observation of individuals not taking part in the training session. Here, the objective is to implement a person-independent application and standalone framework that can be deployed on various platforms.

This paper deals with person-independent posed facial expression recognition (FER) and proposes an automatic FER framework based on the shape and appearance descriptions. Two popular approaches have been proposed in the literature for decoding facial expressions. The first is geometric-based feature extraction. This approach relies on encoding geometric information such as the position, distance, and angle on the facial landmark points that should be first identified by a landmark detector, and then extracts the feature vectors. The second approach is the appearance-based technique, which characterizes the appearance textural information resulting from the facial movements related to each of the emotion classes. Therefore, a set of features is extracted and is expected to contain relevant discriminative information to classify the different classes. The appearance-based approach utilizes many techniques for feature extraction, including those based on different transforms such as wavelet sub-bands, Gabor filters, an optimal matrix factorization and a steerable pyramid transform, an independent component analysis (ICA), a Zernike moments method, a global Gabor-Zernike feature descriptor, a principal components analysis (PCA), and a linear discriminant analysis (LDA) based Fisherface method. Introduced by Ojala's study, local binary patterns (LBPs) constitute a new philosophy in feature extraction. The motivations behind this philosophy rely on overcoming the limitations of global features by allowing the extraction of local relevant features based on pixel neighborhood thresholding and then combining the obtained vectors to construct the final image descriptor. The LBP operator was originally proposed for texture classification, although considering its discriminative power and low computational cost, it has also been adopted in many other computer vision applications, mainly face-related. Since Ojala's study, many state-of-the-art LBP variants have been proposed to enhance the original LBP capabilities. Indeed, researchers are still searching for robust local descriptors with a high discriminative power, and numerous powerful LBP variants continue to be developed in the literature. Notable recent methods include a local optimal oriented pattern (LOOP) [3], local neighborhood difference pattern (LNDP) [42], and local directional ternary pattern (LDTP) [16]. This paper introduces a new automatic FER framework based on a hybrid approach that combines geometric and appearance concepts by extracting the textural and shape features from facial landmarks. The proposed combination is expected to promote an enhanced performance for person-independent FER because we consider geometric and appearance information that carries sufficient relevant features to describe the emotional classes. The geometric representation is obtained by interpolating the positions of 49 keypoints (landmarks) detected in the input image generating a binary patch, which is exploited to compute the shape features using the HOG method. By contrast, the appearance description is also extracted based on the detected landmarks instead of the whole face image, which makes our proposed FER framework able to fulfill the person-independent constraint. The appearance features are extracted from 32 pixel ×32 pixel sub-images centered on each landmark using a brand new handcrafted descriptor, which is referred to as orthogonal and parallel-based directions–generic query map binary patterns (OPD-GQMBP), which is also proposed in this paper. The OPD-GQMBP handcrafted descriptor is based on orthogonality and parallelism geometries for selecting the most prominent neighbors. It adopts an $n \times n$ neighborhood region to extract four feature maps based on four defined thresholding structures for each central pixel. The four feature maps are then decoded into one histogram. Afterwards, the 49 feature vectors are concatenated to form the final appearance feature. During the classification step, we use the SVC library preprocessed by the PCA technique intended to reduce the dimensionality of the feature vectors. The framework is evaluated on five of the most widely used state-of-the-art benchmarks: KDEF, CK+, RaFD, JAFFE, and OuluCasia. To ensure the person-independent evaluation, the leave-one-subject-out (LOSO) protocol is adopted on all five datasets. To prove the superiority of the proposed OPD-GQMBP descriptor for the FER, we conducted a comprehensive experimental comparison against 12 recent and powerful LBP variants and 10 state-of-the-art deep features (e.g., VGG, ResNet, and Inception) using our FER framework on the five datasets. We also compared our FER framework perfor-

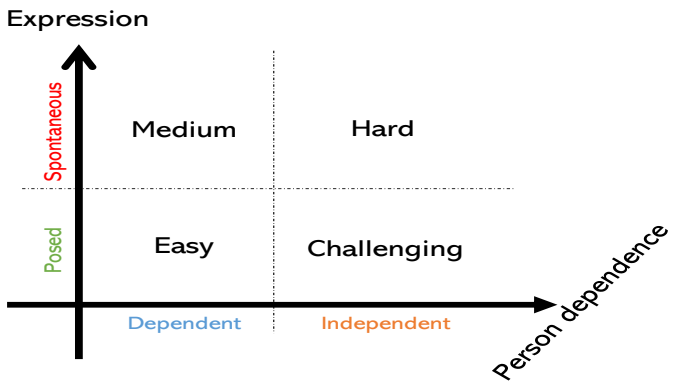

Figure 1: Facial Expression Recognition difficulty levels.

mance to those reported in the literature (from articles in journals with a high impact factor and from highly indexed conferences). The major contributions of our study can be summarized as follows:

- An FER system based on the coupling shape and appearance information is proposed. This system applies the Dlib package to detect 49 landmarks on the facial image and then extracts the shape and textural features before proceeding to the SVM-based classification stage.
- New handcrafted LBP variants our applied for a texture analysis based on the orthogonality and parallelism geometry concepts, which are referred to as OPD-GQMBP. This is a generic descriptor that takes the neighborhood size as a parameter. This parameter offers the opportunity to balance the computational speed with the discrimination performance.
- We considered five widely used benchmarks from the literature for evaluation: CK+, KDEF, JAFFE, Oulu Casia, and RaFD. We adopted an experimental person-independent evaluation on each dataset, using the LOSO protocol.
- The proposed FER framework has outperformed many recent studies published in journals with a high impact factor, including deep-learning and handcrafted frameworks, following the LOSO protocol.
- The performance of the OPD-GQMBP descriptor is fairly evaluated based on its comparison to 12 recent handcrafted LBP variants and 10 deep features. The results showed the superiority of OPD-GQMBP against all of the tested handcrafted and deep features.

To provide readers and field-interested researchers with a better reading experience, this paper is organized as follows. Section 2 presents some state-of-the-art approaches devoted to the FER problem, covering handcrafted and deep-based features. Section 3 introduces the proposed OPD-GQMBP descriptor and highlights its process to compute the textural feature vector. Moreover, this section shows the overall proposed FER framework and explains how the proposed OPD-GQMBP and HOG operators are applied and combined to obtain the textural and shape features from the detected landmarks of a given input face image. Section 4 provides comprehensive experiments on five widely used databases and shows a comparative evaluation of state-of-the-art descriptors (handcrafted-based and deep-based) and FER systems. The last section provide some concluding remarks regarding our study and areas of future research.

2. Related works

The computer vision community has conducted many studies devoted to facial expression recognition (FER) by applying machine learning techniques. In this section, we briefly present some state-of-the-art FER frameworks to highlight some of the proposed architectures that rely on either handcrafted descriptors or deep-learning methods. Shan et al. [31] proposed an approach, referred to as Boosted-LBP, based on combining a basic LBP descriptor with the Adaboost algorithm to enhance the classification performance. They conducted experiments on CK+, MMI, and JAFFE databases, and found that the Boosted-LBP outperforms the basic LBP combined with a multi-class SVM classifier. Moreover, they reported that local methods (LBP) perform better than global methods (Gabor filters). Zhang et al. [48] proposed a novel facial expression recognition method using a local binary pattern (LBP) and local phase quantization (LPQ) based on a Gabor face image. First, Gabor wavelets are applied to capture the prominent visual attributes, which are separable and robust to illumination changes, by extracting multi-scale and multi-direction spatial frequency features from the face image. Then, the LBP and LPQ features based on the Gabor wavelet transform are fused for face representation. Considering that the dimensions of a fused feature are too large, the PCA-LDA algorithm is used to extract compressed features. Finally, the method is tested and verified using multi-class SVM classifiers. Lekdioui et al. [20] proposed an automatic FER framework based on a local appearance approach, extracting the features from seven regions of interest (ROIs) covering the left eyebrow, right eyebrow, left eye, right eye, eyebrows, nose, and mouth. They evaluated the LBP, LTP, and CLBP texture descriptors and their combination with the HOG operator cascading with a linear SVM classifier. They found that the concatenation of LTP and HOG leads to the best FER performance on three datasets (CK, FEED, and KDEF). Their framework strengths rely on extracting the appearance features from seven sub-images defined from landmarks carrying information about the facial expression class, in addition to combining the LBP-Like descriptor with the HOG operator. However, this architecture presents certain drawbacks that we can point out. The seven extracted sub-images have different sizes and orientations, but their computed features have the same length. We found that the nose region of interest is vertically oriented compared to the eyebrow regions, which are horizontal, and the eye regions, which are almost square. Therefore, different amounts of information on different locations are represented over feature vectors of the same length. Furthermore, this study did not cover an important number of handcrafted methods, and no deep-learning method was evaluated. The method proposed by Makhmudkhujaev et al. [22] uses a new handcrafted LBP descriptor referred to as local prominent directional pattern (LPDP) for FER application. It is also an appearance-based approach exploring the benefits of extracting features from three patches: edge, curved edge, and corner-like texture maps. Their study focuses only on the handcrafted descriptor LPDP and its scheme to extract textural features. The authors also used a thresholding parameter to discriminate significant features from insignificant patterns in feature-less/smooth regions of a face. Afterwards, a feature selection method is applied to reduce the dimensionality of the final feature vector because they use the spatial division on the input image. However, this system takes as input the entire face image, which makes it inconvenient for person-independent FER

applications. In addition, no shape descriptor has been adopted in the overall framework, relying only on LPDP extracted features. Minchul et al. [33] used a convolutional neural network model to achieve facial expression recognition. They adopted and aligned cropped faces from FER-2013, SFEW2.0, CK+, KDEF, and Jaffe with respect to the landmark position of the eyes. The training data were augmented 10 times by flipping them. Five types of data input (raw, histogram equalization, isotropic smoothing, diffusion-based normalization, and difference of Gaussian) were tested. They then selected the one that showed the highest accuracy as a target structure for fine-parameter tuning. Finally, the CNN network with histogram equalization images was chosen as the baseline CNN model for further research. Yu et al. [45] proposed a method that contains a face detection module based on an ensemble of three state-of-the-art face detectors, JDA, DCNN, and MoT. Subsequently, a classification module composed of an ensemble of deep convolutional neural networks (CNNs) was adopted based on averaging the output responses. Each CNN model is initialized randomly and pretrained on the Facial Expression Recognition (FER) Challenge 2013 database. The pretrained models were then fine-tuned on the training set of SFEW 2.0. To combine multiple CNN models, they presented two schemes for learning the ensemble weights of the network responses: minimizing the log-likelihood loss and minimizing the hinge losses. According to the results reported in their study, the hinge loss performs slightly better than the log-like and single CNN models on the validation and test sets of the FER2013 and SEFW databases. Therefore, their framework is computationally heavy, and the outcomes are not very promising. Jung et al. [13] proposed a new CNN framework based on combining the temporal appearance and temporal geometry extracted from two CNN models. The faces in the input image sequences are detected, cropped, and rescaled to a pixel resolution of 64×64, and 49 landmark points are then extracted using the IntraFace algorithm. Finally, these two models are combined using an element-wise sum of the outputs of the last fully connected layers from the two temporal CNN models. Through several experiments conducted on the CK+, MMI, and Oulu-CASIA databases as well as numerous data from various data augmentation techniques, the framework built showed that the two models cooperate with each other. However, the joint model did not improve the recognition of all of the facial expressions and achieved the same performance as the temporal appearance and temporal geometry models of the Disgusted, Fear, Happy, and Surprised classes. In addition, the temporal appearance CNN model outperformed the geometry model on all tested databases. Most of the previous methods have considered the entire facial region as the input information, and have paid less attention to the sub-regions of human faces, which may lead to a large difference between the extracted and expected representations. Indeed, when the extracted information obtained from the entire face image is irrelevant, the final recognition result will be affected. Because the judgment of the facial expression is usually based on the information of several sensitive components in some areas of the face, such as the eyes, nose, and mouth, in this paper, a new method is proposed that concentrates the feature extraction on these sub-regions, which not only allows for the extraction of more relevant features, but will also further improve the overall recognition rate.

3. Proposed framework

In this paper, we propose an enhanced framework for facial expression recognition (FER). The system is based on the SVM classifier to predict the class of a given input image. It considers 49 landmark points and extracts and combines the shape and appearance features to be fed to the classifier. To clearly highlight the contributions of our study and describe the workflow of our system in detail, we first describe the new textural handcrafted descriptor referred to as OPD-GQMBP. We then present how it is combined with the HOG shape descriptor to build the overall workflow.

3.1. OPD-GQMBP: New handcarfted descriptor for FER

As discussed in the introduction, the LBP operator is extremely flexible and many of its aspects can be employed to develop enhanced descriptors for specific tasks. In our case, we propose a new LBP variant, referred to as OPD-GQMBP, which is based on new neighborhood topologies leading to four discriminant feature maps, and adopts the LBP original kernel function that outputs low computational codes. The motivation behind the OPD-GQMBP descriptor relies on selecting orthogonal and parallel neighboring pixels that are believed to present the most information within a sub-block. In mathematics, orthogonality is defined as the generalization of the perpendicularity notion, which was adopted by [1], who proposed a reduced LBP version referred to as OC-LBP, which considers two sets of four pixels located on the orthogonal lines. Thus, it produces only a feature histogram with a 2×2^3 feature histogram. . The OPD-GQMBP descriptor is generic and adjustable depending on the needs of the considered application. It adopts a $n \times n$ sub-block neighborhood, where n is an odd integer (3,5,7,9,...), to maintain symmetric neighborhoods. The concept behind this is the selection of prominent pixels within this neighborhood. Given a central pixel I_c, as can be seen in Figure 2, we define four pixel groups, each of which contains $n \times 2$ pixels forming two lines. Two sampling groups are based on orthogonality $\{SG_1{}^{Ort}, SG_2{}^{Ort}\}$, whereas the two others are based on the concept of parallelism, i.e., $\{SG_1{}^{Par}, SG_2{}^{Par}\}$. Therefore, each sampling group SG is defined on two lines $SG_k{}^t(I_c) = \{L^t{}_{k,1}(I_c), L^t{}_{k,2}(I_c)\}$ where t stands for the type (Ort/Par) of the sampling group and k the group number $(1/2)$. Figure 3 shows a Cartezian coordinate system centered on the central pixel I_c to encode the position of each pixel considered in each sampling group within a 7×7 neighborhood. The sampling groups are defined as follows:

$$SG_1{}^{Ort}(I_c) = \begin{bmatrix} L^{Ort}{}_{1,1}(I_c) = \{(x,0)/x \epsilon T\} \\ L^{Ort}{}_{1,2}(I_c) = \{(0,y)/y \epsilon T,\} \end{bmatrix} \quad (1)$$

$$SG_2{}^{Ort}(I_c) = \begin{bmatrix} L^{Ort}{}_{2,1}(I_c) = \{(x,x)/x \epsilon T\} \\ L^{Ort}{}_{2,2}(I_c) = \{(x,-x)/x \epsilon T\} \end{bmatrix} \quad (2)$$

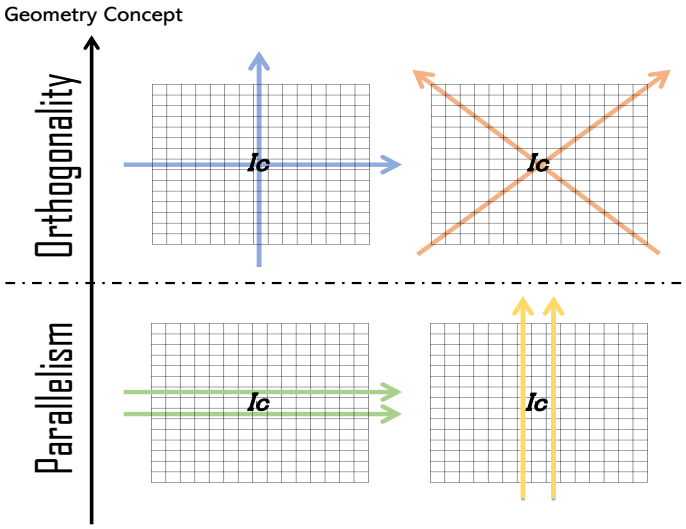

Figure 2: OPD-GQMBP neighborhood topologies.

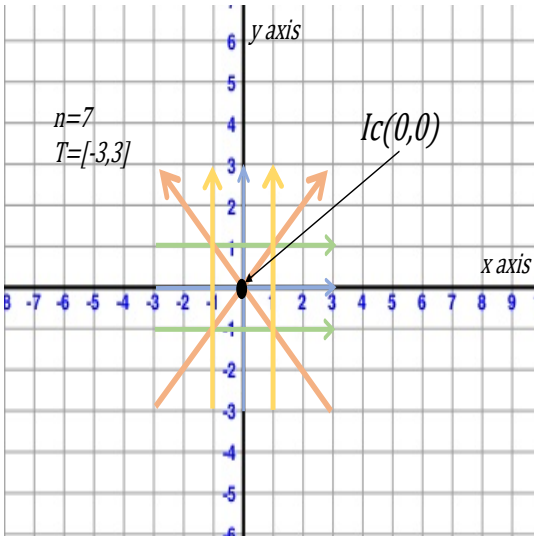

Figure 3: Cartesian system used to identify the pixel coordinates for each line and group.

$$SG_1{}^{Par}(I_c) = \begin{bmatrix} L^{Par}{}_{1,1}(I_c) = \{(x,1)/x\epsilon T\} \\ L^{Par}{}_{1,2}(I_c) = \{(x,-1)/x\epsilon T,\} \end{bmatrix} \quad (3)$$

$$SG_2{}^{Par}(I_c) = \begin{bmatrix} L^{Par}{}_{2,1}(I_c) = \{(1,y)/y\epsilon T\} \\ L^{Par}{}_{2,2}(I_c) = \{(-1,y)/y\epsilon T\} \end{bmatrix} \quad (4)$$

where

$$T = \left[-\tfrac{n-1}{2}, \tfrac{n-1}{2}\right] \quad (5)$$

Here, T is the interval of values defining the coordinates (x, y) of the pixels constructing the two lines $\{L^t{}_{k,1}, L^t{}_{k,2}\}$ of each sampling group $SG_k{}^t(I_c)$.

Because all pixels within the $n \times n$ neighborhood are identified, we can proceed to the thresholding process. In this step, we generate for each sampling group $SG_k{}^t$ its feature map $\Im_k{}^t$, and obtain four feature maps:

$$\Im(I_c) = \begin{Bmatrix} \Im_1{}^{Ort}(I_c) = \Gamma(SG_1{}^{Ort}(I_c)) \\ \Im_2{}^{Ort}(I_c) = \Gamma(SG_2{}^{Ort}(I_c)) \\ \Im_1{}^{Par}(I_c) = \Gamma(SG_1{}^{Par}(I_c)) \\ \Im_2{}^{Par}(I_c) = \Gamma(SG_2{}^{Par}(I_c)) \end{Bmatrix} \quad (6)$$

with

$$\Gamma(SG_k{}^t)(I_c) = \Delta(L^t{}_{k,1}(I_c), L^t{}_{k,2}(I_c)) \quad (7)$$

where Δ is the Heaviside function, which was originally used in the LBP operator defined in Eq 8, and applied the two lines of the same group to the threshold element by element. Thus, the length of the generated binary code for each feature map is the size (n) of the neighborhood, and the number of possible produced patterns is 2^n. Thus, by concatenating the patterns produced by all feature maps, we generate 4×2^n possible patterns. After encoding each pixel in the input image and obtaining the four feature maps, we transform them into a histogram vector as the final descriptor for the image, as defined in

$$\Delta(x,y) = \begin{cases} 1 & , x \geq y \\ 0 & , x < y \end{cases} \quad (8)$$

$$H(F) = \left\langle H^{\Im_1{}^{Ort}}, H^{\Im_2{}^{Ort}}, H^{\Im_1{}^{Par}}, H^{\Im_2{}^{Par}} \right\rangle \quad (9)$$

where

$$H^{\Im_k{}^t}(\mathbf{p}) = \sum_{\chi \in F} \delta(\Im_k{}^t(\chi), \mathbf{p}) \quad (10)$$

In Eq 10, $\mathbf{p} \in [0, 2^n - 1]$ is a pattern used to compare to the patterns $\Im_k{}^t(\chi)$, χ is the gray-scale value of the computed feature image F, and the delta function $\delta(\cdot)$, which is defined as follows (see. Eq. 11):

$$\delta(\mathbf{a}, \mathbf{b}) = \begin{cases} 1, & \text{if } \mathbf{a} = \mathbf{b}; \\ 0, & \text{otherwise} \end{cases} \quad (11)$$

To include more spatial information into the OPD-GQMBP descriptor, the feature image is spatially divided into $w\times$ w small non-overlapping blocks B_i. Then, all corresponding histograms $H(B_i)$ extracted from all blocks are concatenated to form the final holistic image representation through Eq. 12.

$$\mathbb{H} = \prod_{i=1}^{w^2} H(B_i) \quad (12)$$

where $\mathbb{H}$ is the final descriptor, $\prod$ is the concatenation operation, and $H(B_i)$ is the histogram of the OPD-GQMBP descriptor computed on the i^{th} block. Note that each elementary histogram $H(B_i)$ has a length of 4×2^n, whereas the dimensionality of $\mathbb{H}$ is $4 \times 2^n \times w^2$.

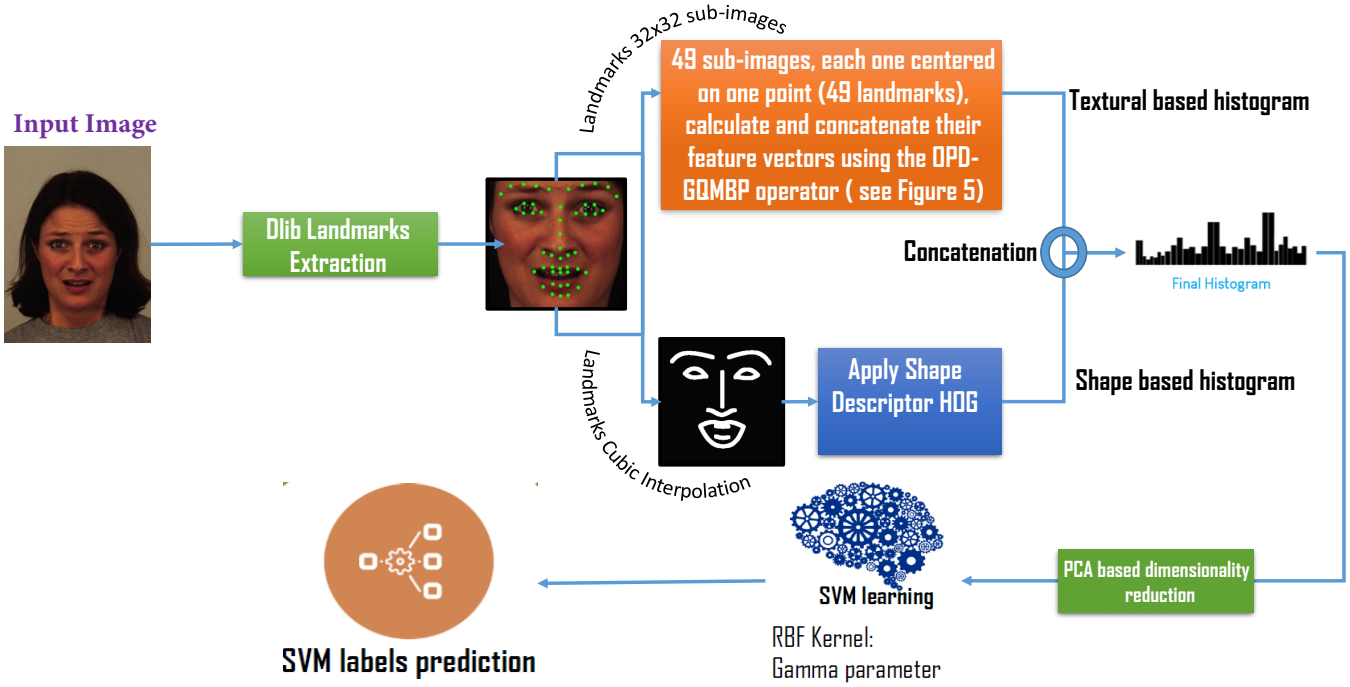

Figure 4: Overall view of the proposed FER framework.

3.2. FER system

After defining the neighborhood topology and thresholding kernel of the OPD-GQMBP descriptor, we now present the overall view of our proposed system for the FER task. The idea behind this framework is to combine the shape and appearance information to provide a more accurate FER, whereas most of the state-of-the-art proposed FER systems rely only on one piece of information, either geometric or appearance based. To do so, we computed the textural and shape features based on the location of 49 detected keypoints (landmarks) on the input face image. The shape representation is obtained by interpolating the 49 landmarks to form curves to be further analyzed using the HOG operator, whereas the appearance representation is based on texture analysis by applying the proposed OPD-GQMBP operator on specific sub-images of the input image. To make the FER system more able to fulfill the person-independent constraint, the appearance features are extracted from sub-images with a pixel resolution of 32×32 and centered on each landmark carrying sufficient and relevant information about the expressed emotion and less irrelevant information of the person's face. Figure 4 illustrates the overall pipeline of the proposed FER system. First, the input image is fed to the dlib landmarks extractor to locate the 49 points (green color). These locations are then interpolated to generate a binary patch of the expressed emotion, upon which the HOG operator is applied to compute the shape feature vector. Meanwhile, the OPD-GQMBP descriptor (or state-of-the-art descriptors for comparison) was used to extract the textural features from each 32×32 sized sub-image centered on one landmark leading to a set of 49 histograms (49 landmarks) that are further concatenated together to construct the appearance feature vector. Note that spatial division was adopted to compute the OPD-GQMBP feature vector by dividing each sub-image into non-overlapping blocks of size $w \times w$, as illustrated in Figure 5. The number of spatial blocks that divide the sub-image depends on the considered dataset and is related to the camera resolution and image blur. Indeed, blurred images require fewer blocks than clear images, which present more details to be detected. At the end of the feature extraction stage, the HOG and OPD-GQMBP computed histogram vectors are concatenated to compose the final image descriptor that is further fed to a dimensionality reduction using the PCA method before proceeding to the classification phase based on an SVM. We used the LIBLINEAR 2.30 library as a multiclass kernel-based vector machine implementation for MATLAB/Python environments. This library provides many classification and regression solvers. We chose the support vector classification based on the Cramer and Singer solver ($Kernel = 4$) as a simplified multi-class SVM. Furthermore, this kernel allows optimized training and takes less time compared to the LibSVM library implementation.

4. Experimental Analysis and Discussions

The previous sections introduced a new framework for FER based on combining the shape and appearance features computed using the HOG and our proposed OPD-GQMBP descriptor. Our FER framework relies on extracting the features from 49 landmark points detected using the Dlib algorithm on each input image, which are believed to carry relevant and sufficient information to recognize the emotion expressed in the input image. To show its effectiveness for the person-independent FER, our system is extensively evaluated on five well-known and widely used benchmarks in the literature: KDEF, CK+, RaFD, JAFFE, and OuluCasia. To ensure person independence in our

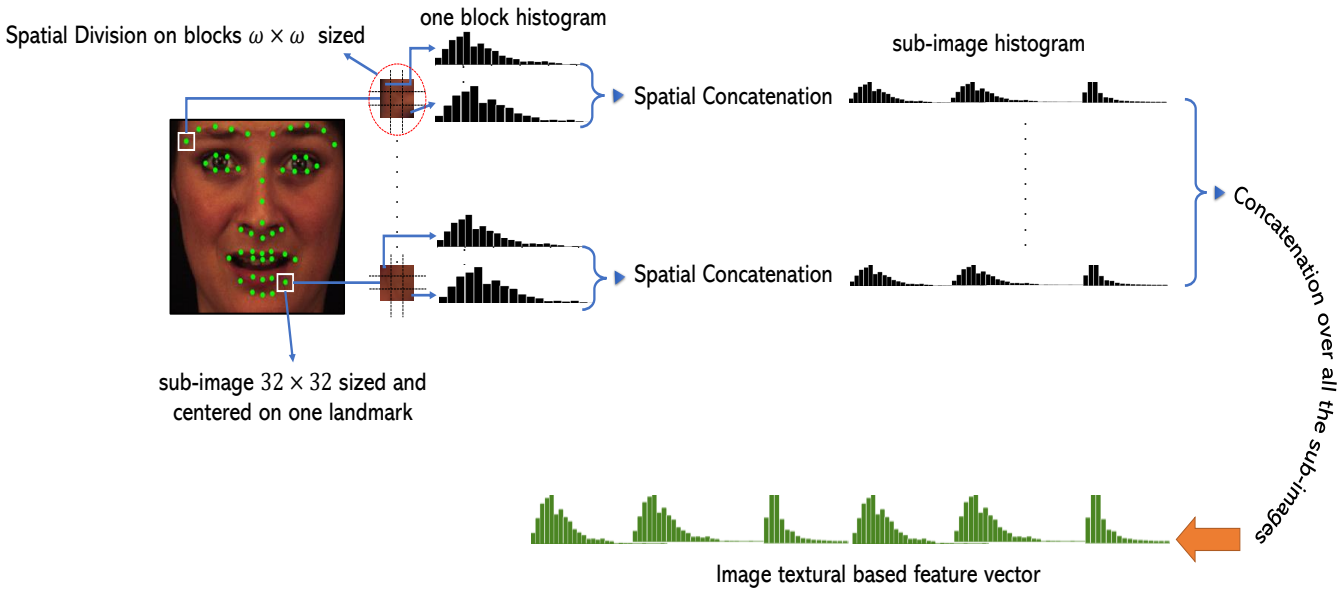

Figure 5: Texture feature extraction workflow based on the proposed OPD-GQMBP operator.

testing, we set up a LOSO protocol, where all samples of one person were excluded from the training set and used for testing. The process is repeated for N-persons, and no prior person information is included in the training stage. As discussed in Section 3, our paper presents two main contributions: the FER framework itself and the OPD-GQMBP handcrafted descriptor. To highlight the results of each of the contributions, we first evaluated four possible configurations of the OPD-GQMBP descriptor, and then compared its performance against 12 recent handcrafted methods and 10 state-of-the-art deep features (see Table 1) within our FER framework, keeping the same evaluation protocol and conditions. Afterwards, the performance of the proposed FER framework is compared to those presented in previous state-of-the-art purchase, published in highly indexed and well-known journals.

4.1. Experimental datasets

- The Japanese Female Facial Expression (JAFFE) dataset has 213 facial expression images, representing the 7 basic emotions: Anger (30 images), Disgust (29 images), Fear (32 images), Happiness (31 images), Sadness (31 images), Surprise (30 images), and Neutral (30 images). Figure 6 illustrates some examples of facial expressions. This database is extremely challenging regarding the particularity of Japanese females that have similar face features, generating more inter-class visual features

- The Karolinska Directed Emotional Faces (KDEF) is a widely used dataset for evaluating FER methods. It includes 70 individuals (50% men, 50% women) that uniformly express basic emotions over two sessions, leading to a total of 980 images. In our experiments, we considered only one session to have only one observation per emotion for each person (490 images), which resulted in

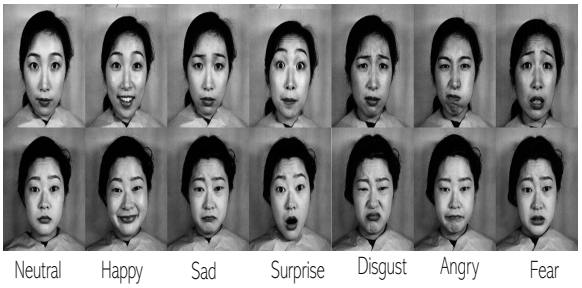

Figure 6: Samples of two subjects from JAFFE database

70 samples per class. Figure 7 shows the observation of each class for a female from this database.

Figure 7: Samples of a subject from KDEF database

- The Cohn-Kanade v2 database (CK+) is a sequence-based database. It contains 593 image sequences from 123 subjects. The first image of each sequence represents the neutral state of the subject, whereas the peak of the emo-

tion is represented at the end of the sequence. During our experiment, we selected only the last frame to construct the sets of six emotions, whereas the neutral class is constructed by the first frame from each sequence. The Angry class has 45 samples; Disgust, 59; Fear, 25; Happy, 69; Neutral, 45; Sad, 28; and Surprise, 82. Some of these samples are shown in Figure 8.

Figure 8: The seven emotions of one person from CK+ database

- The Oulu-CASIA NIR and VIS expression database is also a sequence-based dataset, including 80 subjects (South Asian and Caucasian) with the six typical expressions. The videos are recorded using two imaging systems: near infrared (NIR) and visible light (VIS). Only the last frame from each sequence of VIS in the database is considered, and the neutral expression is represented by the first frame. Therefore, a dataset of 560 images was obtained (80 samples per class). As can be seen in Figure 9, the images are slightly blurry and unclear, and the visual features of the South Asian individuals are quite similar, making this a challenging dataset.

Figure 9: Samples of two subjects from the OuluCasia database

- The Radboud Faces Database (RaFD) is composed of 67 individuals (including Caucasian, Moroccan, and Dutch adults, and Caucasian children, both boys and girls) displaying 8 emotional expressions. In addition to the seven basic emotion expressions, this database includes the Ccontempt facial expression, which can be similar to angry and disgust emotions, but also expresses the feeling of dislike for and superiority over another person, and/or his that person's actions. Moreover, the Ccontempt emotion is not symmetric and occurs only on one side of the face.

Figure 10 displays shows the 8 facial expressions of a person from the RaFD database.

Figure 10: Samples of one subject from RaFD database

4.2. *Deep-Features for Facial Expression Recognition*

In this study, the proposed descriptor is compared to the 10 deep features described in the literature. Deep methods are inspired and rely mainly on convolutional neural networks (CNNs). Owing to the technological progress made in the GPU computational field, a CNN has proven to be one of the most widely used techniques for computer vision applications, such as image classification, object detection, and face recognition. A CNN typically consists of convolutional layers, pooling layers, and fully connected layers. Convolutional layers are the core building blocks of a CNN.

We considered 10 deep networks, which are briefly introduced in the following, based on a recent survey [19]:

- AlexNet: Referring to its author Alex Krizhevsky, AlexNet was proposed in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC 2012). AlexNet is a deeper configuration of a LeNet5 network. Therefore, the high performance at this competition comes at the cost of a high computation that was possible using only graphic card units. It consists of five convolutional layers, two fully connected hidden layers, and one fully connected output layer.

- VGG: VGG network architectures were introduced by Simonyan and Zisserman in 2014. VGG stands for the Visual Geometry Group of Oxford University. Compared to LeNet and AlexNet, VGG networks are conceptually simple employing only stacked 3×3 convolutional layers combined with a max pooling layer to reduce the volume size, leading to two fully connected layers of 4096 nodes each, followed by a softmax classifier. VGG19 has three more convolutional layers than VGG16.

- ResNet: Residual learning networks were also proposed for the ILSVRC competition in 2015, introducing the Skip Connection concept to CNNs, which are known as recurrent networks. Typical ResNet models are implemented with double- or triple-layer skips that contain nonlinearities (ReLU) and batch normalization between them. The skip connection technique allows the training of 152 layers or more with fewer computations then AlexNet and VGG networks. In this study, we considered ResNet18, ResNet50, and ResNet101.

- DenseNet: Densely connected convolutional networks were inspired by the ResNet topology. They incorporate dense residual blocks composed of batch normalization, ReLU activation, and a 3×3 convolution. The ResNet models use the sum function as a skip connection, whereas DenseNet integrates the concatenation. Therefore, each input layer receives all outputs of the earlier versions. The concatenation process generates an output with a large number of channels, which makes DenseNet models computationally heavy.

- Inception: Google proposed its own deep learning inspired by LeNet, referred to as Inception, stacking more convolutional layers deeper to achieve a better performance, which comes at the cost of heavy computations and a complex design. The philosophy of inception relies on concatenating the responses of different convolution filters at the same layer, forming the input of the next layer. Moreover, they used a 1×1 convolution filter as a feature reduction technique before jumping to the next layer. Google introduced four versions of the Inception architecture, Inception.v1 known as GoogLeNet with 27 layers, Inception.v2, Inception.v3, and Inception.v4 tackling batch normalization, factorization, and grid size control problems, respectively. Google proposed two versions of a residual network inspired by the performance of ResNet, known as InceptionResNet.v1 and Inception-ResNet.v2 based on creating the skip connections on the previous Inception models. Inception-ResNet.v1 and Inception-ResNet.v2 networks have the same computational cost of Inception.v3 and Inception.v4, respectively.

To employ deep learning architectures for deep feature extraction in solving the FER problem, we follow the basic procedure shown in Figure 11. Initially, the model was trained end-to-end on a big dataset, mainly the Facial Expression Recognition 2013 (FER2013) database. Afterwards, the model is expected to achieve a good training performance using the validation set. We then proceed to the transfer learning technique to extract the features of the subject database that belongs to the same application as the database used for the initial training (same classes). Once the features are obtained, we train the SVM classifier and evaluate the performance of each deep feature. To further improve this deep-based FER architecture, the calculated features may be concatenated with other descriptors such as HOG, Gabor filters, and LBP descriptors before proceeding to the classification step.

4.3. Evaluation of OPD-GQMBP neighborhood size configuration

The proposed OPD-GQMBP descriptor is a generic method defined by the neighborhood size n, which can be seen as a user-specified parameter depending on the needs of the considered application. To find the best value for FER, we conducted an experiment evaluating the performance of four configurations: n = 3, 5, 7, and 9. For each, we evaluated the FER framework on the five datasets using the LOSO protocol. The smaller neighborhood sizes provide less computational cost, but with weak discriminative power, and larger sizes enhance the discriminative power, but require more resources to store and classify the extracted features. For example, a neighborhood size of 3 ($OPD{-}GQMBP^3$) generates only $4{\times}2^3 = 32$ possible patterns, whereas neighborhood sizes of 5 ($OPD{-}GQMBP^5$), 7 ($OPD{-}GQMBP^7$), and 9 ($OPD - GQMBP^9$) produce 128, 512, and 2048 patterns, respectively. Table 2 shows the recognition rates obtained from this experiment. It can be concluded that with a higher neighborhood size, we obtain more discriminative feature extraction. The most effective configuration is n=7, which managed to reach the top performance on 4 databases. Thus, the 512 generated patterns proved to be sufficient for characterizing the seven emotional classes. Here, $OPD - GQMBP^9$ achieved a top accuracy of 97.53% on the CK+ database outperforming $OPD - GQMBP^7$, but suffered a performance drop on the other datasets. Indeed, in some cases, methods that generate a high number of patterns may cause a performance drop owing to pattern redundancy. It is clear that with $OPD - GQMBP^3$ the configuration cannot outperform the other configurations; however, the recorded accuracies remained prominent regarding the low computation (32 patterns only). The performances of $OPD - GQMBP^3$ and $OPD - GQMBP^5$ are extremely similar with small variations. We acknowledge that we could not evaluate neighborhood sizes of greater than 9 owing to the required computational resources (out of memory). Therefore, we adopt $OPD - GQMBP^7$ in the rest of this paper because it corresponds to the best among the tested configurations.

4.4. Comparative analysis against state-of-the-art handcrafted and deep feature methods

The goal of this comprehensive analysis is to compare the performance of the proposed OPD-GQMBP descriptor to those recorded in the literature on feature extraction methods, including handcrafted and deep-based approaches. We considered 12 recent and powerful LBP variants published in high-impact journals and proposed for various applications, mainly texture classification and face recognition. In addition, we evaluated the top-10 state-of-the-art deep learning models proposed thus far. These models were initially trained on the FER2013 database, each of them reached a validation accuracy of above 60% on 25,000 images, which can be considered extremely significant. We then use transfer learning to extract the features of the five databases adopted in this study. We respected the same evaluation protocol (LOSO) to provide a fair and systematic analysis. Table 3 lists the performance achieved by each method or model. We provide two metrics. The first metric is the average accuracies recorded for all runs of each database depending on the number of individuals, of which JAFFE has 10 runs, CK+ has 106, KDEF has 70, OuluCasia has 80, and the RaFD database has 67. The second metric is the maximum accuracy reached for all runs per database. We highlight the top 3 average values with green color.

It can be seen from the average accuracies that the proposed OPD-GQMBP descriptor managed to score the top performance on all tested datasets. For the CK+ database, the

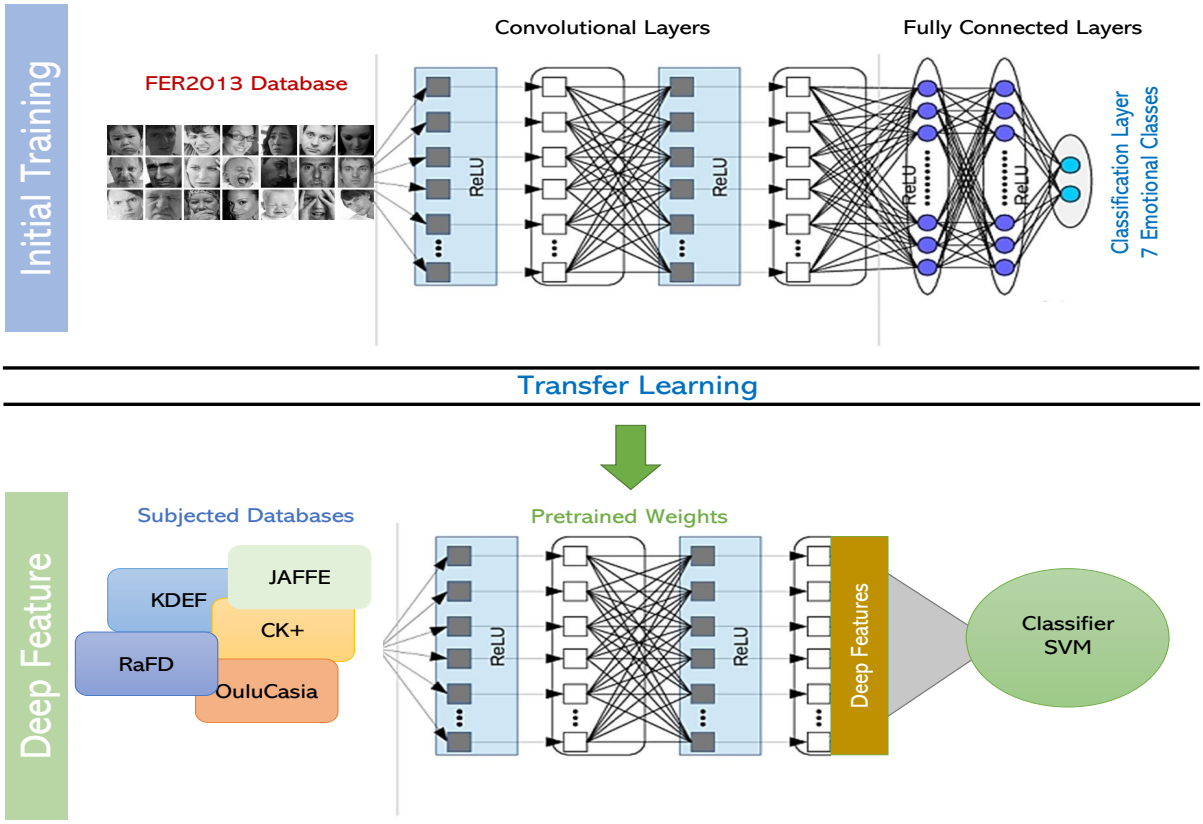

Figure 11: Deep-feature based FER approach.

OPD-GQMBP descriptor achieved a score of 96.48% on 106 runs with a maximum recognition accuracy of 100%, maintaining more than a 2% gap over the next best method, i.e., the handcrafted MNTCDP descriptor, which managed to secure an average accuracy of 94.36%. All of the evaluated methods were capable of reaching a maximum accuracy of 100% on the CK+ dataset. Moreover, the majority of the handcrafted methods performed at above 87.81% when applying the DCP descriptor, whereas the deep features reached a maximum average of 88.76% with VGG19 only and 54.18% (AlexNet) as the minimum average for CK+. As indicated in Table2, we found that n=9 for the configuration of the OPD-GQMBP method scored 97.53% on CK+, which was a 1% improvement compared to OPD-GQMBP with n=7 adopted in this evaluation. In addition, OPD-GQMBP reached 77.62% on JAFFE with a lead of 3% over the rest of the methods, with the second highest average accuracy being 74.76% for the DCP descriptor followed by ELGS as the third-best performing method at 73.81%. According to the maximum accuracy on the JAFFE database, only the OPD-GQMBP and DCP methods managed to score 100%. The Japanese females with high facial similarities make the JAFFE an extremely tough benchmark for FER systems. The results demonstrate that the JAFFE database is quite challenging, particularly for deep features where the top average accuracy was limited to 56.67%, as obtained by VGG16, whereas the handcrafted features granted an average accuracy of above 64.76%, which was reached by the LOOP descriptor. For the KDEF database, the first remark to be concluded is that the OPD-GQMBP descriptor is the only method that breaks the 90% average performance ceiling, and the best state-of-the-art method reached 88.37% (LDTP). The performance of the handcrafted methods and deep methods varied from 78.57% to 88.37% and from 65.31% to 76.73%, respectively. The lead gap on this database was again approximately 2% in favor of the OPD-GQMBP descriptor, which proves its discriminative power for FER applications. Despite the quality of the recorded images in terms of resolution, lighting conditions, and a uniform background, the KDEF database is also challenging in face of CK+. It presents fewer individuals (fewer runs) expressing the seven emotions in different manners, resulting in more intraclass similarities and approaching spontaneous facial expressions. OuluCasia can be considered as the toughest among the adopted benchmarks owing to the blur images and the South Asian individuals composing this database. The OPD-GQMBP descriptor scored 77.32% as the overall best performing method, followed by ARCS-LBP (75.5%) and DC (74.64%) descriptors. The lowest accuracy was recorded by the AlexNet deep network, reaching only 29.64%. In addition, all methods achieved a 100% maximum accuracy except for MNTCDP, AlexNet, GoogLeNet, and Inceptionv3. Moreover, only the ELGS, DC, ARCS-LBP, and OPD-GQMBP methods exceeded a 70% average accuracy, whereas the DCP descriptor, which had the second highest rank on JAFFE, found its performance limited to 45.89% as the lowest average accuracy over all handcrafted LBP-like methods.

Table 1: Summary of the Handcrafted and Deep feature methods of the literature tested and evaluated in this paper

N°	Abbreviation	Complete name (Application)	Year
1	DSLGS [5]	Difference Symmetric Local Graph Structure (Finger vein recognition)	2015
2	DCP [4]	Multi-Directional Multi-Level Dual-OPD-GQMBPs Patterns (Face recognition)	2015
3	DRLBP [23]	Dominant Rotated Local Binary Patterns (Texture classification)	2016
4	QBP [47]	Quad Binary Pattern (Target tracking)	2016
5	ELGS [9]	Extended Local Graph Structure (Texture classification)	2016
6	LNDP [28]	Local neighborhood difference pattern (Texture classification)	2017
7	DC [42]	Rotation-invariant features based on directional coding (Texture classification)	2018
8	LCCMSP [8]	Local Concave-and-Convex Micro-Structure Patterns (Texture classification)	2018
9	LDTP [16]	Local directional ternary pattern (Texture classification)	2018
10	LOOP [3]	Local Optimal Oriented Pattern (Spieces recognition)	2018
11	MNTCDP [14]	Mixed Neighborhood Topology OPD-GQMBPs Decoded Patterns (Face recognition)	2018
12	ARCS-LBP [9]	Attractive-and-Repulsive Center-Symmetric Local Binary Patterns (Texture classification)	2019
13	VGG16	Visual Geometry Group - University of Oxford - 19 Layers	2014
14	VGG19	Visual Geometry Group - University of Oxford - 16 Layers	2014
15	ResNet18	Residual Neural Network - Microsoft Research - 18 Layers	2015
16	ResNet50	Residual Neural Network - Microsoft Research - 50 Layers	2015
17	ResNet101	Residual Neural Network - Microsoft Research - 101 Layers	2015
18	AlexNet	Convolutional Neural Network by Alex Krizhevsky - 8 Layers	2012
19	DenseNet	Densly Connected CNN - Cornell Tsinghua Univs & Facebook - 201 Layers	2017
20	GoogLeNet	Inception version1 based on LeNet - Google - 22 Layers	2015
21	Inception v3	Inception version3 - Google - 159 Layers	2015
22	InceptionResNet v2	Inception-ResNet-version2 CNN (Xception) - Google - 126 Layers	2017

Table 2: Average FER rate of each OPD-GQMBP configuration (neighborhood size) for all databases

N_{Size} Config	CK+	JAFFE	KDEF	OuluCasia	RaFD
$OPD-GQMBP^3$	95.74	73.33	88.57	73.93	96.08
$OPD-GQMBP^5$	96.01	73.33	87.96	74.82	95.9
$OPD-GQMBP^7$	96.48	78.57	90.2	77.32	97.39
$OPD-GQMBP^9$	97.53	71.9	87.96	75.95	96.08

The RaFD database is collected by a set of well-trained individuals that clearly express eight emotions (basic emotions + contempt). Indeed, the peak average accuracy exceeded 97% by the proposed OPD-GQMBP descriptor with a minor lead (0.19%) against the LDTP and LCCMSP methods (97.01%). Moreover, all of the handcrafted descriptors reached above 90%, except for DCP (87.31%). By contrast, ResNet50 was the best among the deep feature methods, with an average accuracy of 84.51%. In terms of stability, the OPD-GQMBP descriptor performed well on the five datasets, always reaching the top average accuracies and 100% at all times. The ELGS method also presented a stable performance across all databases. By contrast, DCP suffered a performance decrease on the KDEF and OuluCasia datasets. For deep feature methods, VGG16 can be considered as the best performing deep feature method.

The deep learning networks did not perform well on the five benchmarks, despite reaching a validation accuracy of above 60% on 25,000 images of the FER2013 database. The problem here is that the deep learning methods should be fine-tuned on each dataset before extracting the features to obtain satisfactory results. The applied application in this paper is person-independent, and to ensure that the probe images of a given person are unseen by the framework, we should perform fine tuning on each run and for each deep method. Hence, because we have a set of 333 persons on the five datasets and we consider 10 deep learning models, we need 3330 fine tunes to expect satisfying results from these models, which are time- and resource-consuming. Moreover, such frameworks are intended for real implementations and deployments, and fine tuning is not always a possible option. In addition, the probe images will have generally different characteristics than the trained images.

4.5. Comparison against state-of-the-art FER systems

In this subsection, we compare the results obtained by our proposed FER framework to those achieved by previous studies in the field of facial expression recognition. Tables 4, 5, 6, 7 and 8 list the highest accuracies on the five datasets reported in well-indexed journals and conferences of the literature. We tried to collect the maximum number of studies that followed the same adopted evaluation protocol (person-independent).

As can be seen in Table 4, the proposed FER framework outperforms all listed systems, including both handcrafted and deep-based features in the CK+ database. We reached an accuracy of 96.48% (97.53% with the OPD-GQMBP at a neighborhood size of n=9), whereas the best accuracy of the other state-of-the art methods was 94.96% achieved using LPQ with an SVM classifier. Moreover, the majority of the published approaches reached between 90% and 94%. In the JAFFE database, the accuracies of the other state-of-the-art methods are low compared to that of CK+, where the maximum reported accuracy is 76.46% scored by the CFER-based framework, which was outperformed by our FER framework (77.62%). The proposed framework managed to surpass with a significant margin all other approaches on the KDEF dataset, except for the WCFN (89.55%) and AlexNet (89.33%) based systems, where the margin is small (0.65% and 0.87% of our framework performance,

Table 3: Average and maximum accuracies recorded on the five datasets by each method.

	Method	JAFFE		KDEF		CK+		OuluCasia		RaFD	
		Avg	Max	Avg	Max	Avg	Max	Avg	Max	Avg	Max
Deep Features	VGG16	56.67	76.19	76.53	100	86.34	100	56.96	100	82.84	100
	VGG19	53.81	76.19	76.73	100	88.76	100	58.57	100	81.53	100
	ResNet18	35.24	57.14	72.24	100	72.67	100	52.32	100	83.02	100
	ResNet50	44.76	61.9	75.71	100	77.32	100	56.96	100	84.51	100
	ResNet101	52.86	71.43	70.82	100	76.79	100	53.57	100	78.17	100
	AlexNet	43.81	66.67	66.94	100	54.18	100	29.64	71.43	67.54	100
	DenseNet	48.1	80.95	70.2	100	76.15	100	50.36	100	77.8	100
	GoogLeNet	47.14	66.67	71.63	100	69.98	100	44.64	85.71	79.66	100
	Inceptionv3	39.52	57.14	65.31	100	71.22	100	44.46	85.71	70.52	100
	InceptionResNetv2	47.62	66.67	76.33	100	80.17	100	55	100	81.72	100
Handcrafted LBP Variants	ELGS	73.81	95.24	85.71	100	92.59	100	70	100	95.34	100
	DSLGS	60.95	85.71	78.57	100	91.2	100	64.64	100	91.23	100
	MNTCDP	69.05	85.71	85.51	100	94.36	100	55.36	85.71	95.15	100
	QBP	62.38	85.71	79.39	100	91.99	100	66.43	100	91.79	100
	DRLBP	70	85.71	84.49	100	90.35	100	65.36	100	96.08	100
	LNDP	69.52	95.24	86.94	100	92.86	100	68.75	100	96.08	100
	DCP	74.76	100	69.18	100	87.81	100	45.89	100	87.31	100
	DC	66.67	95.24	84.69	100	89.87	100	74.64	100	95.52	100
	LCCMSP	71.43	95.24	86.12	100	92.28	100	69.29	100	97.01	100
	LOOP	64.76	85.71	85.92	100	91.7	100	69.64	100	96.27	100
	LDTP	70	90.48	88.37	100	93.3	100	67.5	100	97.01	100
	ARCS-LBP	65.24	95.24	85.51	100	91.45	100	75.5	100	96.64	100
	Proposed OPD-GQMBP descriptor	**77.62**	**100**	**90.2**	100	**96.48**	100	**77.32**	100	**97.2**	100

Table 4: State-of-the-art person-independent FER accuracies on CK+ database

Methods	Type	Avg accuracy
LPDP [22]	Handcrafted	94.5
DCNN [27]	Deep	94.44
DNN [26]	Deep	93.52
CNN+AFM [43]	Deep	89.84
AlexNet+SVM [43]	Deep	86.83
GoogLeNet [43]	Deep	85.71
STM-ExpLet [21]	Deep	94.13
LTeP+SVM [2]	Handcrafted	94.93
LPQ+SLPM+NN [41]	Handcrafted	94.61
WPLBP [6]	Handcrafted	91.72
Proposed	Handcrafted	**97.53**

Table 5: State-of-the-art person-independent FER accuracies on JAFFE database

Methods	Type	Avg accuracy
LTeP+SVM [2]	Handcrafted	67.14
LPQ+SLPM+NN [41]	Handcrafted	67.61
EDR-PCANet [39]	Deep	69.4
C-classLDA-NN [18]	Deep	74.73
LBP based LDA [32]	Handcrafted	73.4
CFER [38]	Handcrafted	76.46
Features fusion [29]	Handcrafted	70
Proposed	Handcrafted	**77.62**

Table 6: State-of-the-art person-independent FER accuracies on KDEF database

Methods	Type	Avg accuracy
AlexNet+FC6+LDA [10]	Deep	89.33
HOG+SRC [34]	Handcrafted	78
VGG-Face Deep [46]	Deep	72.55
SCAE [30]	Deep	86.73
DFD [40]	Handcrafted	82.24
WCFN [44]	Deep	89.55
MobileNet [11]	Deep	73.74
EDR-PCANet [39]	Deep	80.61
Proposed	Handcrafted	**90.2**

Table 7: State-of-the-art person-independent FER accuracies on OuluCasia database

Methods	Type	Avg accuracy
STM-ExpLet [21]	Deep	74.59
LBP+Gabor+SVM [50]	Handcrafted	74.37
HOG 3D [17]	Handcrafted	70.63
AdaLBP [49]	Handcrafted	73.54
Atlases [11]	Deep	75.52
Proposed	Handcrafted	**77.32**

Table 8: State-of-the-art person-independent FER accuracies on RaFD database

Methods	Type	Avg accuracy
Visual Attention CNN [37]	Deep	95.2
DS+FE+GEM+SVM [25]	Handcrafted	90.8
LPQ+FE+GEM+SVM [25]	Handcrafted	94.4
LBP+FE+GEM+SVM [25]	Handcrafted	94.5
Metric Learning [12]	Deep	95.95
BAE-BNN-3 [35]	Deep	96.93
W-CR-AFM [43]	Deep	96.27
Net1-Net2 [36]	Deep	93.41
Proposed	Handcrafted	**97.2**

respectively). On the OuluCasia dataset, the proposed framework (77.32%) outperformed all other state-of-the-art methods, where the top accuracy was limited to 75.52%, which was achieved by Atlases. On the RaFD database, our proposed framework obtained 97.2%. Many studies described in the literature achieved an accuracy of nearly 97%. However, the majority applied only seven emotion classes. Overall, we conclude that the proposed facial expression recognition framework managed to outperform all of the state-of-the-art methods tested.

4.6. Confusion matrix-based analysis for the FER

The confusion matrix allows the performance of the recognition according to each label (each emotion in our case). Through this chart, we are capable of analyzing the recognition rate of each emotion, as well as which are the easiest and most difficult emotions to recognize. In addition, this analysis allows us to identify which emotions affect the others. Figures 12, 13, 14, 15 and 16illustrate the confusion charts generated from the results of our proposed FER framework for CK+, JAFFE, KDEF, OuluCasia, and RaFD, respectively.

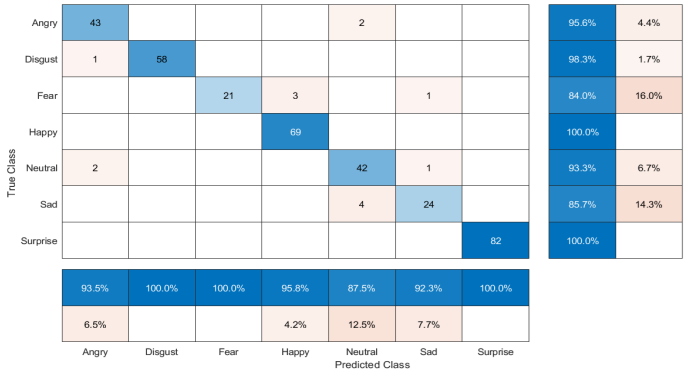

Figure 12: Confusion matrix of the seven emotions of CK+

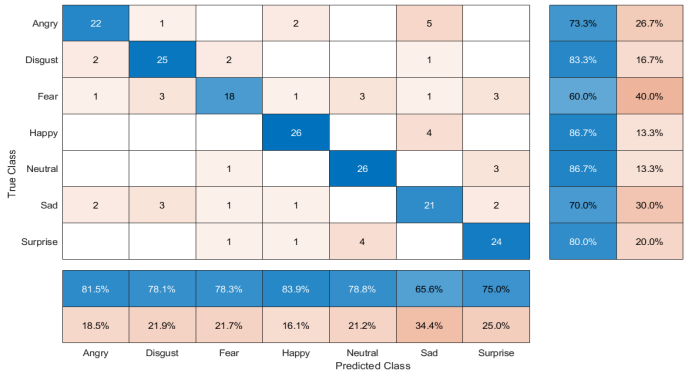

Figure 13: Confusion matrix of the seven emotions of JAFFE

In the CK+ database, the Happy and Surprise classes were perfectly recognized, whereas Fear and Sad experienced the highest misclassification rate (16.0% and 14.3%, respectively). Fear was confused three times with Happy and once with Sad. A Neutral emotion was the most affective emotion (with a 12.5% false-negative rate) and was predicted four times in the case of

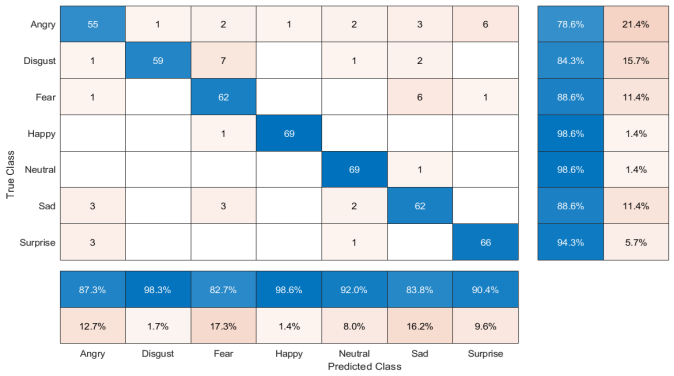

Figure 14: Confusion matrix of the seven emotions of KDEF

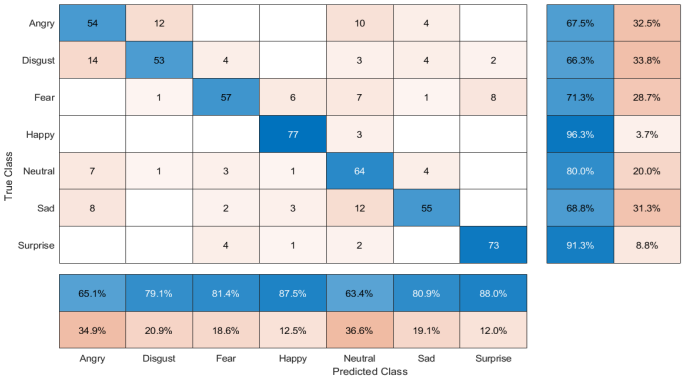

Figure 15: Confusion matrix of the seven emotions of OuluCasia

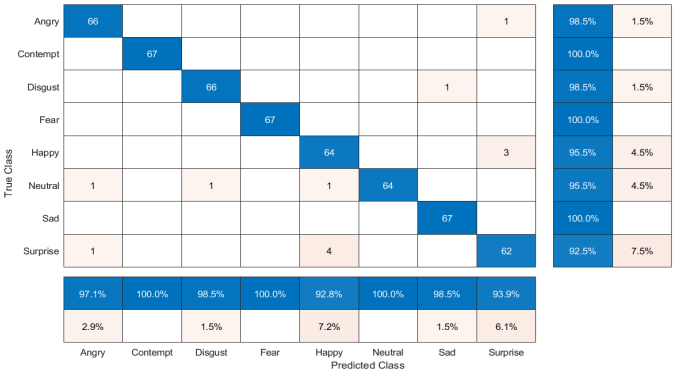

Figure 16: Confusion matrix of the eight emotions of RaFD

Sad and twice for Angry. For the JAFFE database, Happy and Neutral were the easiest to identify with an accuracy of 86.7%, and the hardest was Fear (60%), which was confused with all emotions, i.e., three times with Disgust, Neutral, and Surprise, and once with the remaining emotions. The Sad class presented the highest false-negative rate (34.4%). The females composing this database express Angry and Sad emotions in similar ways because five Angry samples were identified as Sad. For the KDEF dataset, the errors were less than those of JAFFE. Happy and Neutral once again were highly recognized emotions (98.6%), but Angry was a challenging class, with a rate of only 78%. Our FER framework confused Disgust seven times with Fear and confused Fear six times with Sad. Happy and Disgust did not affect any other emotion except once each for Angry. As expected from OuluCasia, the misclassification errors were extremely high. The Happiness emotion is the only one to be highly recognized with an accuracy of 96.3%, followed by Surprise (91.3%). The remaining accuracies were between 66.3% and 80%. In addition, Angry and Neutral dramatically influenced the other emotions by 34.5% and 36.6%, respectively. In the RaFD database, all rates were high with perfect recognition rates for three classes (i.e., Contempt, Fear, and Sad) in addition to Angry and Disgust, which were misclassified only once. However, there is a mutual confusion between Happy and Surprise because three Happy samples were identified as Surprise and Surprise samples were identified as Happy four times. For all databases, we determined that Happy and Neutral are the most recognized emotions, and that Neutral is an extremely perturbing framework, presenting high false-negative rates on many benchmarks.

4.7. Implementation and execution time

The FER experiments presented were conducted on an Alienware Aurora R8 with a 4.6 GHz Core i7-8th Processor Boost, 12 threads, and 48 GB of RAM, running with the Ubuntu 18.04.2 LTS (Bionic Beaver) operating system and equipped with two GTX1080Ti GPUs. The developed framework is coded using Python 3.7 and MATLAB 2019b environments. The authors are willing to share all codes used by the community once this paper is published. The computational cost is one of the key performance indicators considered in machine learning applications. Therefore, we ran an experiment that calculates the elapsed time required to predict the label of a given input image with a pixel resolution of 762×562, highlighting the execution time of each step of the proposed framework:

- Dlib landmark detection.
- Shape feature extraction based on HOG descriptor.
- Appearance feature extraction using the proposed OPD-GQMBP as well as the state-of-the-art handcrafted methods denoted as ”getFeatures”.
- PCA dimensionality reduction.
- Label prediction using the SVM library.

We excluded the deep-learning methods from this evaluation because they require GPUs to perform the feature extraction; thus, it will be unfair to compare CPU-based methods with GPU-based approaches regarding the computation power of the GPUs. The CPU can be used to calculate the feature vector using a deep model, but it takes approximately 4 s to compute it for an input image with a size of only 224×224 (three-times smaller than the size of the original image), which is extremely high compared to the handcrafted features. The obtained computational times are included in Table 9 and illustrated graphically in Figure 17 (for greater readability). As can be seen, the elapsed times for the landmark detection, HOG shape feature

extraction, and SVM prediction did not change across the evaluated methods, where they recorded times of 15.9, 19.58, and 2.306 ms, respectively. The process of extracting the appearance features was demonstrated to be the most time-consuming within our framework (requiring more than 50% of the total time). The fastest handcrafted method is DRLBP, which extracted the features from the 49 landmarks in 31.5 ms, whereas LDTP took 311.6 ms to extract these features and is thus judged as the heaviest approach. However, the DRLBP did not perform well in terms of classification accuracy, whereas the proposed OPD-GQMBP descriptor managed to offer an execution time of only 90.34 ms, which is quite beneficial regarding its high performance as demonstrated earlier. Moreover, all of the best-performing descriptors took more than 100 ms to compute the appearance features. Although the PCA stage is common to all descriptors for a dimension reduction, its execution time was variable and affected by the number of generated patterns of each handcrafted method as a high number of patterns leads to more computations. Nevertheless, it can be remarked that the PCA computation times are similar for the methods sharing the same sized patterns generated. The methods producing 256 patterns such as LOOP, QBP, DC, ARCS-LBP, and DRLBP recorded a PCA computation time of approximately 11 ms. The PCA process of the proposed OPD-GQMBP descriptor as well as those producing 512 patterns took approximately 20 ms, whereas LDTP and LNDP methods generating 1024 patterns took approximately 36 ms. By contrast, the PCA-based dimensionality reduction process related to the LCCMSP (2048 patterns) descriptor was completed within 83.5 ms. Overall, we can disclose that the proposed framework along with the OPD-GQMBP method managed to predict the label of an image with a pixel resolution of 762×562 in less than 150 ms, allowing a processing of 7 frames per second, which is considered as a real-time feedback according to the specifications of person-independent FER systems [7, 24].

Table 9: Elapsed time in milliseconds to compute the features and predict the label of an input image

Method	Dlib	Hog	getFeatures	PCA	SVM	Total
ARCS-LBP	15.9	19.58	61.4	11.6	2.235	110.715
DC	15.9	19.58	167.3	10.1	2.235	215.115
DCP	15.9	19.58	62.37	20.4	2.235	120.485
DRLBP	15.9	19.58	31.6	11.7	2.235	81.015
DSLGS	15.9	19.58	30.8	20.6	2.235	89.115
ELGS	15.9	19.58	65.22	20.6	2.235	123.535
LCCMSP	15.9	19.58	145.72	83.5	2.235	266.935
LDTP	15.9	19.58	311.6	36.21	2.235	385.525
LNDP	15.9	19.58	34.22	35.5	2.235	107.435
LOOP	15.9	19.58	293.2	11.7	2.235	342.615
MNTCDP	15.9	19.58	101.8	20.5	2.235	160.015
OPD-GQMBP	15.9	19.58	90.34	20	2.235	148.055
QBP	15.9	19.58	35.5	12.4	2.235	85.615

5. Conclusion and Future works

In this paper, we proposed a new handcrafted descriptor, OPD-GQMBP, along with a framework for facial expression recognition. We considered using individual people as the most challenging task acquiring the intention of many computer vision researchers, as can be seen from the number of studies published in this context. The OPD-GQMBP method applies orthogonality and parallelism geometrics to cover the most prominent pixels within a neighborhood. OPD-GQMBP is a generic descriptor with the neighborhood size as a parameter, which can be adjusted to obtain the most discriminative feature extraction regarding the intended application. For FER, we found that the neighborhood size of $n = 7$ is the best configuration of the proposed OPD-GQMBP descriptor. With regard to the developed FER framework, we combined the shape and texture feature extraction techniques applied considering 49 landmarks detected by the Dlib package on each input image. This strategy allowed a powerful FER system to outperform many existing methods in the literature on five widely used benchmarks. To fairly judge the performance of the OPD-GQMBP descriptor, a comparative analysis was conducted on five benchmarks. We assessed the performance of 12 handcrafted and 10 state-of-the-art deep feature methods applied within the developed framework and respecting the same evaluation protocol. Based on the analysis of the experimental results, it was inferred that the OPD-GQMBP descriptor managed to outperform all other methods, including deep features, and reached the top recognition rates. The accuracies recorded by the developed FER system combined with the proposed OPD-GQMBP descriptor were the highest in comparison to the those reported by the existing state-of-the-art FER systems that adopted the same protocol (person-independent LOSO). Although it is true that our system managed to outperform many of the state-of-the-art systems, it needs improvement on databases containing Asian individuals (e.g., JAFFE and Oulu CASIA) because they tend to present similar facial features, which confuse the classifier in the case of the LOSO protocol. We believe that the ultimate solution is to reduce the number of extracted appearance features and focus on the binary patch calculated based on the detected landmarks. This patch should incorporate more information, and not only the landmark location, and therefore handcraft a shape descriptor to obtain the most prominent information. Moreover, this proposal will help in the development of generic person-independent FER systems because the input images will be coded into a common patch. Even though our framework, along with the proposed descriptor, performs efficiently in terms of computational time, we think that there is room for improvement using other dimension reduction strategies and/or landmark selections. We are also investigating an enhancement of the performance of our framework by utilizing other sophisticated classifiers than an SVM and combining learnable features with the proposed OPD-GQMBP descriptor. In addition, we intend to extend the set of the studied emotion classes with the compound classes, reaching 20 different classes. We also considered creating a mixed database from existing databases to gather all challenges in a unique benchmark, which will offer more challenging testing and evaluation.

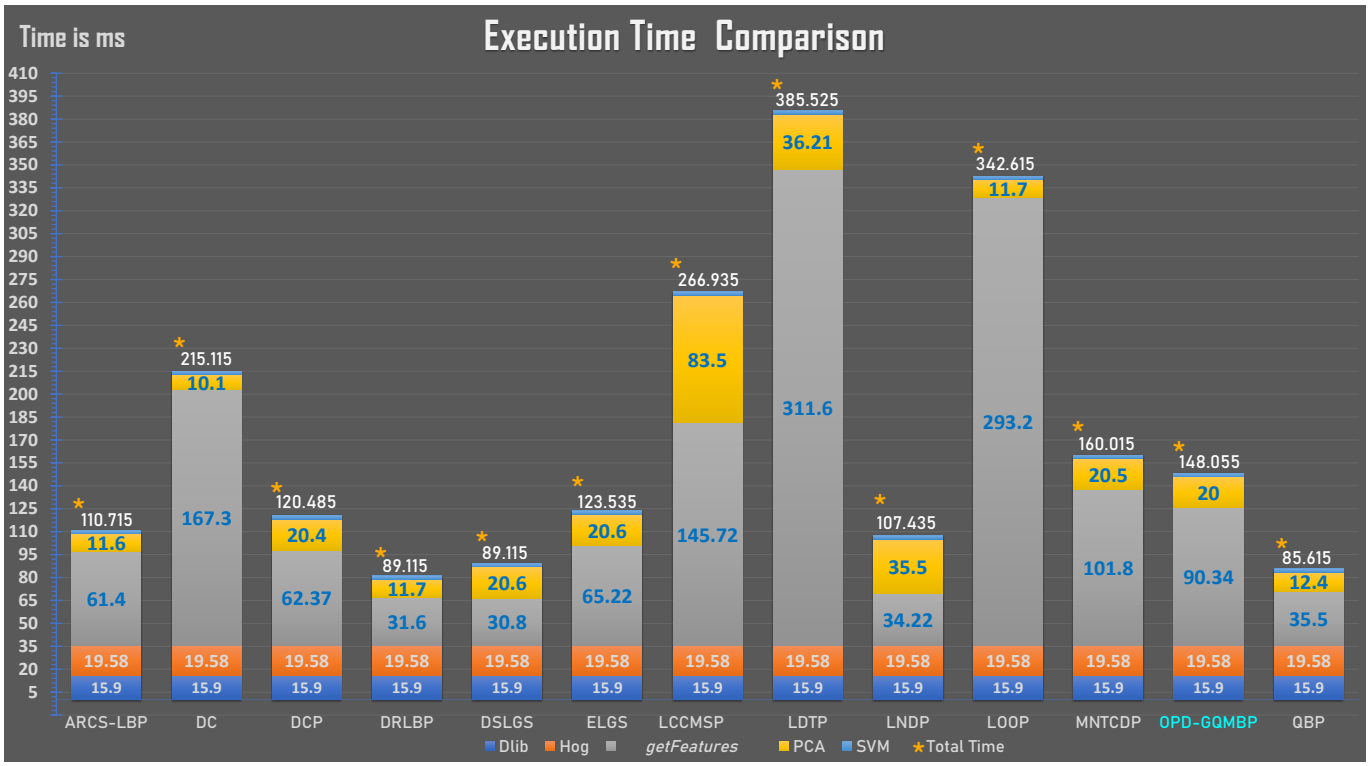

Figure 17: Elapsed time in milliseconds required to compute the features and predict the label of an input image

6. Acknowledgments

The authors gratefully acknowledge the funding received from CNSRT-Maroc (Centre National de la Recherche Scientifique et Technique) and the French government (Eiffel scholarship).

[1] Barkan, O., Weill, J., Wolf, L., and Aronowitz, H. (2013). Fast high dimensional vector multiplication face recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 1960–1967.

[2] Bashar, F., Khan, A., Ahmed, F., and Kabir, M. H. (2014). Robust facial expression recognition based on median ternary pattern (mtp). In *2013 International Conference on Electrical Information and Communication technology (EICT)*, pages 1–5. IEEE.

[3] Chakraborti, T., McCane, B., Mills, S., and Pal, U. (2018). Loop descriptor: Local optimal-oriented pattern. *IEEE Signal Processing Letters*, 25(5):635–639.

[4] Ding, C., Choi, J., Tao, D., and Davis, L. S. (2016). Multi-directional multi-level dual-cross patterns for robust face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(3):518–531.

[5] Dong, S., Yang, J., Wang, C., Chen, Y., and Sun, D. (2005). A new finger vein recognition method based on the difference symmetric local graph structure (dslgs). *International Journal of Signal Processing, Image Processing and Pattern Recognition*, 8(10):71–80.

[6] Du, L. and Hu, H. (2019). Weighted patch-based manifold regularization dictionary pair learning model for facial expression recognition using iterative optimization classification strategy. *Computer Vision and Image Understanding*, 186:13–24.

[7] Duncan, D., Shine, G., and English, C. (2016). Facial emotion recognition in real time. *Stanford University*.

[8] El merabet, Y. and Ruichek, Y. (2017). Local concave-and-convex micro-structure patterns for texture classification. *Pattern Recognition*.

[9] Elmerabet, Y. et al. (2019). Attractive-and-repulsive center-symmetric local binary patterns for texture classification. *Engineering Applications of Artificial Intelligence*, 78:158–172.

[10] Fei, Z., Yang, E., Li, D. D.-U., Butler, S., Ijomah, W., Li, X., and Zhou, H. (2020). Deep convolution network based emotion analysis towards mental health care. *Neurocomputing*.

[11] Guo, S., Feng, L., Feng, Z.-B., Li, Y.-H., Wang, Y., Liu, S.-L., and Qiao, H. (2019). Multi-view laplacian least squares for human emotion recognition. *Neurocomputing*, 370:78–87.

[12] Jiang, B. and Jia, K. (2016). Robust facial expression recognition algorithm based on local metric learning. *Journal of Electronic Imaging*, 25(1):013022.

[13] Jung, H., Lee, S., Yim, J., Park, S., and Kim, J. (2015). Joint fine-tuning in deep neural networks for facial expression recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 2983–2991.

[14] Kas, M., Ruichek, Y., Messoussi, R., et al. (2018). Mixed neighborhood topology cross decoded patterns for image-based face recognition. *Expert Systems with Applications*, 114:119–142.

[15] Kayyal, M. H. and Russell, J. A. (2013). Language and emotion: certain english–arabic translations are not equivalent. *Journal of Language and Social Psychology*, 32(3):261–271.

[16] Khadiri, I. E., Chahi, A., merabet, Y. E., Ruichek, Y., and Touahni, R. (2018). Local directional ternary pattern: A new texture descriptor for texture classification. *Computer Vision and Image Understanding*, 169:14–27.

[17] Klaser, A., Marszałek, M., and Schmid, C. (2008). A spatio-temporal descriptor based on 3d-gradients.

[18] Kyperountas, M., Tefas, A., and Pitas, I. (2010). Salient feature and reliable classifier selection for facial expression classification. *Pattern Recognition*, 43(3):972–986.

[19] Lateef, F. and Ruichek, Y. (2019). Survey on semantic segmentation using deep learning techniques. *Neurocomputing*, 338:321–348.

[20] Lekdioui, K., Messoussi, R., Ruichek, Y., Chaabi, Y., and Touahni, R. (2017). Facial decomposition for expression recognition using texture/shape descriptors and svm classifier. *Signal Processing: Image Communication*, 58:300–312.

[21] Liu, M., Shan, S., Wang, R., and Chen, X. (2014). Learning expressionlets on spatio-temporal manifold for dynamic facial expression recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1749–1756.

[22] Makhmudkhujaev, F., Abdullah-Al-Wadud, M., Iqbal, M. T. B., Ryu, B., and Chae, O. (2019). Facial expression recognition with local prominent

directional pattern. *Signal Processing: Image Communication*, 74:1–12.

[23] Mehta, R. and Egiazarian, K. (2016). Dominant rotated local binary patterns (drlbp) for texture classification. *Pattern Recognition Letters*, 71:16–22.

[24] Michel, P. and El Kaliouby, R. (2003). Real time facial expression recognition in video using support vector machines. In *Proceedings of the 5th international conference on Multimodal interfaces*, pages 258–264.

[25] Moeini, A. and Moeini, H. (2014). Multimodal facial expression recognition based on 3d face reconstruction from 2d images. In *International Workshop on Face and Facial Expression Recognition from Real World Videos*, pages 46–57. Springer.

[26] Mollahosseini, A., Chan, D., and Mahoor, M. H. (2016). Going deeper in facial expression recognition using deep neural networks. In *2016 IEEE Winter conference on applications of computer vision (WACV)*, pages 1–10. IEEE.

[27] Ouellet, S. (2014). Real-time emotion recognition for gaming using deep convolutional network features. *arXiv preprint arXiv:1408.3750*.

[28] Ouslimani, F., Ouslimani, A., and Ameur, Z. (2018). Rotation-invariant features based on directional coding for texture classification. *Neural Computing and Applications*, pages 1–8.

[29] Poursaberi, A., Noubari, H. A., Gavrilova, M., and Yanushkevich, S. N. (2012). Gauss–laguerre wavelet textural feature fusion with geometrical information for facial expression identification. *EURASIP Journal on Image and Video Processing*, 2012(1):17.

[30] Ruiz-Garcia, A., Elshaw, M., Altahhan, A., and Palade, V. (2017). Stacked deep convolutional auto-encoders for emotion recognition from facial expressions. In *2017 International Joint Conference on Neural Networks (IJCNN)*, pages 1586–1593. IEEE.

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

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

[33] Shin, M., Kim, M., and Kwon, D.-S. (2016). Baseline cnn structure analysis for facial expression recognition. In *2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pages 724–729. IEEE.

[34] St, A. (2017). Emotion recognition: The influence of texture's descriptors on classification accuracy. In *Beyond Databases, Architectures and Structures. Towards Efficient Solutions for Data Analysis and Knowledge Representation: 13th International Conference, BDAS 2017, Ustroń, Poland, May 30-June 2, 2017, Proceedings*, volume 716, page 427. Springer.

[35] Sun, W., Zhao, H., and Jin, Z. (2017a). An efficient unconstrained facial expression recognition algorithm based on stack binarized auto-encoders and binarized neural networks. *Neurocomputing*, 267:385–395.

[36] Sun, W., Zhao, H., and Jin, Z. (2018a). A complementary facial representation extracting method based on deep learning. *Neurocomputing*, 306:246–259.

[37] Sun, W., Zhao, H., and Jin, Z. (2018b). A visual attention based roi detection method for facial expression recognition. *Neurocomputing*, 296:12–22.

[38] Sun, Y. and Wen, G. (2017). Cognitive facial expression recognition with constrained dimensionality reduction. *Neurocomputing*, 230:397–408.

[39] Sun, Z., Hu, Z., and Zhao, M. (2019). Automatically query active features based on pixel-level for facial expression recognition. *IEEE Access*, 7:104630–104641.

[40] Sun, Z., Hu, Z.-P., Wang, M., and Zhao, S.-H. (2017b). Discriminative feature learning-based pixel difference representation for facial expression recognition. *IET Computer Vision*, 11(8):675–682.

[41] Turan, C. and Lam, K.-M. (2018). Histogram-based local descriptors for facial expression recognition (fer): A comprehensive study. *Journal of visual communication and image representation*, 55:331–341.

[42] Verma, M. and Raman, B. (2018). Local neighborhood difference pattern: A new feature descriptor for natural and texture image retrieval. *Multimedia Tools and Applications*, 77(10):11843–11866.

[43] Wu, B.-F. and Lin, C.-H. (2018). Adaptive feature mapping for customizing deep learning based facial expression recognition model. *IEEE access*, 6:12451–12461.

[44] Ye, Y., Zhang, X., Lin, Y., and Wang, H. (2019). Facial expression recognition via region-based convolutional fusion network. *Journal of Visual Communication and Image Representation*, 62:1–11.

[45] Yu, Z. and Zhang, C. (2015). Image based static facial expression recognition with multiple deep network learning. In *Proceedings of the 2015 ACM on international conference on multimodal interaction*, pages 435–442.

[46] Zavarez, M. V., Berriel, R. F., and Oliveira-Santos, T. (2017). Cross-database facial expression recognition based on fine-tuned deep convolutional network. In *2017 30th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, pages 405–412. IEEE.

[47] Zeng, H., Chen, J., Cui, X., Cai, C., and Ma, K.-K. (2016). Quad binary pattern and its application in mean-shift tracking. *Neurocomputing*, 217:3–10.

[48] Zhang, B., Liu, G., and Xie, G. (2016). Facial expression recognition using lbp and lpq based on gabor wavelet transform. In *2016 2nd IEEE International Conference on Computer and Communications (ICCC)*, pages 365–369. IEEE.

[49] Zhao, G. and Pietikainen, M. (2007). Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE transactions on pattern analysis and machine intelligence*, 29(6):915–928.

[50] Zhao, L., Wang, Z., and Zhang, G. (2017). Facial expression recognition from video sequences based on spatial-temporal motion local binary pattern and gabor multiorientation fusion histogram. *Mathematical Problems in Engineering*, 2017.