Multimed Tools Appl (2018) 77:19455–19479
https://doi.org/10.1007/s11042-017-5354-x

Facial-expression recognition based on a low-dimensional temporal feature space

Taoufik Ben Abdallah[1] · Radhouane Guermazi[2] · Mohamed Hammami[3]

Received: 24 January 2017 / Revised: 17 September 2017 / Accepted: 25 October 2017 /
Published online: 16 November 2017

Abstract This paper suggests a facial-expression recognition in accordance with face video sequences based on a newly low-dimensional feature space proposed. Indeed, we extract a Pyramid of uniform Temporal Local Binary Pattern representation, using only XT and YT orthogonal planes (PTLBP^{u2}). Then, a Wrapper method is applied to select the most discriminating sub-regions, and therefore, reduce the feature space that is going to be projected on a low-dimensional feature space by applying the Principal Component Analysis (PCA). Support Vector Machine (SVM) and C4.5 algorithm have been tested for the classification of facial expressions. Experiments conducted on CK+ and MMI, which are the two famous facial-expression databases, have shown the effectiveness of the approach proposed under a lab-controlled environment with more than 97% of recognition rate as well as under an uncontrolled environment with more than 92%.

Keywords Facial-expression recognition · Pyramid of uniform Temporal Local Binary Pattern (PTLBP^{u2}) · Principal Component Analysis (PCA) · Discriminating sub-regions · Low-dimensional feature space

✉ Taoufik Ben Abdallah
taoufik.benabdallah@fsegs.rnu.tn; taoufik.tba@gmail.com

Radhouane Guermazi
r.guermazi@seu.edu.sa

Mohamed Hammami
mohamed.hammami@fss.rnu.tn

[1] MIR@CL Laboratory, Faculty of Economics and Management of Sfax (FSEGS), University of Sfax, Sfax, Tunisia

[2] Saudi Electronic University, Riyadh, Kingdom of Saudi Arabia

[3] MIR@CL Laboratory, Faculty of Sciences of Sfax (FSS), University of Sfax, Sfax, Tunisia

1 Introduction

Automatic Facial-Expression Recognition (AFER) is one of the most active field, and impacts important multimedia applications such as interactive education [19], business office [7, 29], monitoring [7, 19] and Medcine [19]. In distance education i.e., e-learning, the facial-expression recognition is used to evaluate automatically the learner's level of course comprehension, and therefore the online-tutor can adjust the presentation style. In business office, the facial-expression recognition could be a great help to know costumers' impressions useful to analyze the feedback of advertising, marketing, etc. In Monitoring, the facial-expression recognition can control the tiredness of a driver, reveal lies, facilitate social agents, etc. In medicine, the facial-expression recognition can access pain, boost significantly the success of the behavioral therapy, etc.

The most related works [12, 20, 26, 33, 41, 56, 62, 63] have considered the six universal facial expressions formalized by Ekman [11]: surprise, fear, disgust, happiness, sadness and anger. They are generally performed by three main steps: face detection and tracking, feature detection, and machine learning [12]. Facial-expression recognition approaches based on video sequences focus generally on analyzing and detecting only spatial facial features in a one frame in a video sequence i.e., peak frame [25, 26, 28, 37, 60]. Recently, they have considered both spatial and temporal facial features so as to describe the facial movement changes over the entire video sequence [3, 17, 41, 62, 63]. Despite the efficiency of these approaches, the major drawback is the high-dimensional feature space like Zhao et al. [62] that have detected more than 12000 features. Likewise, the majority of related works have considered subject-independent of the facial-expression recognition under a lab-controlled environment. As a matter of fact, the expressions are artificial, and the faces are captured in frontal pose with clear illumination and with the same resolution of frames. Some works [12, 17, 26, 33, 43, 61] have taken into account subject-independent of the facial expression recognition under an uncontrolled environment in which the faces have captured in different conditions i.e., the variability of the expression intensity, the illumination and the resolution. The results obtained are very low compared to those of the facial expression-recognition under a lab-controlled environment. These results have not generally exceeded 70% of recognition rate [12, 61].

Unlike the most related works, the main goal of this paper is to present an approach via using only the temporal discriminating features for facial-expression recognition of subject-independent under an uncontrolled environment. Thus, we have defined a newly-low dimensional feature space, using the discriminating sub-regions of the face. We show that this space can describe sufficiently the facial movement changes. Our approach could also be a competitor to the other literature one like [12] and [17] under a lab-controlled environment as well as under an uncontrolled environment.

The remainder of this paper is organized as follows. Section 2 presents the different approaches for facial-expression recognition in the literature. Section 3 describes the facial-expression recognition approach proposed, and investigates its steps. Section 4 presents experimental results and discusses the effectiveness of our approach. Section 5 offers concluding remarks, and forecasts perspectives.

2 Related works

A survey of facial-expression recognition leads to distinguish two orientations. The first one lies in Static Feature-Based approaches i.e., a spatial description of the features. The second one lies in Dynamic Feature-Based approaches i.e., spatio-temporal description of the features.

The Static Feature-Based approaches can be classified into two main categories of methods: Geometry-Based methods and Appearance-Based methods. The Geometry-Based methods use fiducial points for describing the shape rules (displacements, distances, etc.). These methods have been largely employed by many researchers such as Pu et al. [37], Suk et al. [47], Su et al. [46] and Tian et al. [49]. The Geometry-Based methods have two drawbacks. The first one is the complexity of training classifiers, allowing the localization and tracking of fiducial points. The second one is the imprecision of these classifiers overlooking noise. In fact, the training datasets must contain faces in different poses and lighting, and have to be annotated by a set of fiducial points in order to have a robust classifier that works regardless of constraints. Using Appearance-Based methods can restrict the problems of geometry-feature accuracy by describing the texture changes of the face. One of the most frequently-used features is Gabor filters [32, 40]. These filters have been considered as a very useful tool in computer vision and image analysis due to their optimal-localization properties in both spatial and frequency analysis [8]. Another appearance feature based on bit-planes has been proposed. It consists of eight level-information of bits: bit-plane 0 represents the least significant bits while the last bit-plane 7 shows the most significant bits that provide a binary image having visual resemblance to the original grey-level image. Bit-plane features have been introduced in several areas of researches like face recognition [30, 50], palmprint recognition [30] and facial-expression recognition [51]. Local Binary Pattern (LBP) [39, 42] and Histogram of Oriented Gradient (HOG) [9, 24, 25] have recently aroused increasing interest in image processing and computer vision and shown its effectiveness in a number of applications, in particular for facial-expression recognition. Likewise, LBP has been developed with a large number of variations for improving performance in different applications. There are Centralized Binary Patterns (CBPs) [13], Local Ternary Pattern (LTP) [16], Pyramid of Local Binary Pattern (PLBP) [26], etc. The drawback of Appearance-Based methods is that they produce an extremely large number of non-discriminating features. It is in this context that several studies in the literature have been proposed to select the discriminating features by using AdaBoost [43], PCA [8], mutual information [34], etc. Note that some researchers such as Kotsia et al. [28], Zhang et al. [60] and Wan et al. [54] combine both Geometry-Based methods and Appearance-Based methods for facial-expression recognition.

The Dynamic Feature-Based approaches are focused on detecting the dynamic shape features and/or the dynamic texture features. Dynamic shape features are tracking the displacements of points located on the face between successive frames to capture facial movement changes. Several researchers like Sanchez et al. [41], Shin et al. [45] and Wang et al. [55] have relied on dynamic shape features for facial-expression recognition. Dynamic texture features are sequences of images in scene movements exhibiting certain stationary time properties [10]. Zhao et al. [62, 63] have proposed a variety of dynamic variants of LBP i.e., Volume Local Binary Pattern (VLBP) and LBP on Three Orthogonal Planes (LBP-TOP). They have shown that using LBP-TOP is more efficient than using VLBP. The combination of VLBP and LBP-TOP slightly improves the performance of the classifier, but it increases its complexity (resulting 31176 features) [62]. LBP-TOP has been used by many researchers such as Ji et al. [20], Wang et al. [56] and Yu et al. [59] for characterizing facial appearance changes. Some researchers like Fan et al. [12] and Chen et al. [3] have recently proposed Dynamic Feature-Based approaches based on a combination of spatio-temporal texture and shape features so as to improve facial-expression recognition.

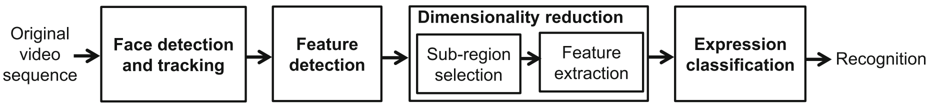

Fig. 1 An overview of the facial-expression recognition approach suggested

3 Methodology

We propose to build an approach for an automatic video facial-expression recognition. It is based on a low-dimensional temporal feature space in order to produce a classifier, making it possible to recognize the facial expression represented by a video sequence that contains only one face of one subject in a frontal pose. Figure 1 shows an overview of the approach proposed. It is performed in four main steps: (i) face detection and tracking, (ii) feature detection, (iii) dimensionality reduction and (iv) expression classification.

3.1 Face detection and tracking

In the approach proposed, we apply the Viola and Jones object detection algorithm [53] on the first frame of video sequences for automatic face detection. This algorithm is based on the Haar appearance features, the Adaboost training and the cascading classifiers. We have selected this algorithm thanks to its efficient and fast features computation, and its robustness in the precision detection of frontally-positioned faces. Then, we carry out the Kanade-Lucas-Tomasi (KLT) algorithm [44] for face tracking across the video frames. KLT algorithm initializes a set of points on the detected face and tracks them, using the differential method for optical flow estimation. We also convert all frames to grayscale, resize them to 64×64 pixels resolution- the most resolution used in the literature [20, 34]- and apply histogram equalization [4] in order to standardize the lighting contrast.

3.2 Feature detection: pyramid of uniform temporal local binary pattern representation (PTLBP^{u2})

Based on the static method proposed in [26] that is performed to detect a pyramidal spatial representation of features, we propose a new feature space where we describe the dynamic-facial movement changes via texture analysis in time. We have called it Pyramid of uniform Temporal Local Binary Pattern representation (PTLBP^{u2}). Precisely, the pyramid is defined as a combination of a set of levels. There, we decompose the frames of a video sequence into sub-regions of different sizes. For each level, we consider only the two temporal planes XT and YT of uniform LBP-TOP (LBP^{u2}-TOP) [62] to detect features, noted TLBP^{u2} features. XT represents the horizontal plane describing the motion of one row in time, and YT represents the vertical plane describing the motion of one column in time. The choice of using only XT and YT temporal planes ("XT+YT") without XY spatial plane to define our feature space has been made referring to a study of the different uses of spatial and temporal planes of LBP^{u2}-TOP: "XT", "YT", "XY+XT", "XY+YT", "XT+YT" and "XY+XT+YT".

Formally, Three main steps can be distinguished:

1. We decompose each video sequence frame into sub-regions according to pyramid levels. Each level l is composed of $r = 4^l$ equal sub-regions. We use only three levels of pyramid- i.e., level 1 ($r = 4$ sub-regions), level 2 ($r = 16$ sub-regions) and level 3 ($r = 64$ sub-regions)- to avoid the increase of the memory space required

(Fig. 2a). We have defined, therefore, four combinations of levels: “level 1 + level 2”, “level 1 + level 3”, “level 2 + level 3”, and “level 1 + level 2 + level 3”. The choice of the best combination of the levels is discussed in off-line experiments (Section 4.3).

2. We calculate the TLBP^{u2} histogram ($H_{l,j}$), for each level l ($l \in 1, 2, 3$), and for each sub-region j ($j \in [1, 2, .., 4^l]$) (Fig. 2b). Firstly, we calculate the coordinates x_p, y_p and t_p) of the P_{pl} neighbor pixels, noted $g_{p\_}pl$, with pl corresponding to temporal XT or YT plane. Bilinear interpolation [15] is applied for determining the pixels that do not figure in each grid of the planes. The number of neighbor pixel for XT and YT planes is equal to 8 i.e., $P_{\mathrm{XT}} = P_{\mathrm{YT}} = 8$, and the distance between center pixel, noted g_c and its neighbor pixels is equal to 1 i.e., $R_{\mathrm{X}} = R_{\mathrm{Y}} = R_{\mathrm{T}} = 1$. The choice of the parameters value is due to the empirical study processed in [62]. Then, we threshold the P_{pl} neighbor pixels of each center pixel g_c. Then, we encode the binary number of

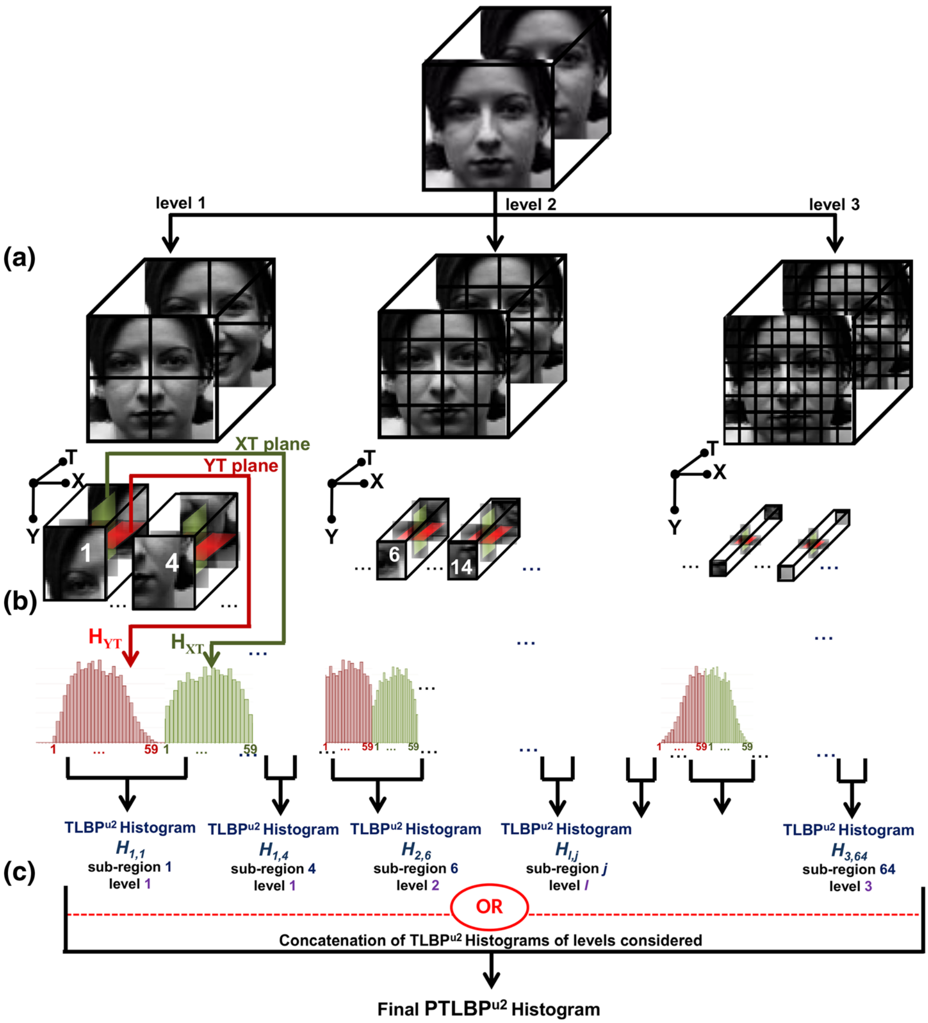

Fig. 2 Process of extracting pyramid of uniform temporal local binary pattern (PTLBP^{u2}) histogram

Algorithm 1 Calculation of the histogram, $H_{pl\_l,j}$

Input : n, m and t, the height, width and number of frames of the video sequence respectively.
$Data_{(n,m,t)}$, the vector of the grey level of all the pixels in video sequence considered.
pl, the plane with pl =XT or pl =YT.
P_{pl}, the number the neighbor pixels of the pl plane.
g_c, the center pixel.
$g_p\_pl(x_p; y_p; t_p)$, the neighbor pixel p of the pl, plane with $p = 1, 2, \ldots, P_{pl}$.
R_{X}, R_{Y} and R_{T}, the radius in axes X, Y and T respectively.

Output: $H_{pl\_l,j}$, the histogram of LBP^{u2} of the pl plane for the sub-region j of level l.

```
begin
    P_pl = 8 ; R_X = R_Y = R_T = 1
    LBP_g_c_pl = 0
    H_pl_l,j = H_pl_l,j[0, 0, ..., 0]_(1,59) /* each column of the vector H_pl corresponding
      to the index (0,1,..,58) of the histogram bin of LBP^u2 on the pl plane */
    for k = 2 to t − 1 do
        for i = 2 to n − 1 do
            for j = 2 to m − 1 do
                g_c = Data(i; j; k)
                for p = 0 to P_pl − 1 do
                    if pl corresponds to the XT plane then
                        g_p_pl = Data(i − R_X sin(2πp/P_pl); j; k − R_T cos(2πp/P_pl))
                    end
                    else
                        g_p_pl = Data(i; j − R_Y cos(2πp/P_pl); k − R_T sin(2πp/P_pl))
                    end
                    if (g_p_pl − g_c) > 0 then
                        LBP_g_c_pl = LBP_g_c_pl + 1
                    end
                end
                if LBP_g_c_pl is uniform then
                    /* the binary pattern decToBinary(LBP_g_c_pl) containing at
                    most two bitwise transitions from 0 to 1 or vice versa, with
                    decToBinary as the function converting a decimal number to
                    the binary number */
                    H_pl_l,j[LBP_g_c_pl + 1] = H_pl_l,j[LBP_g_c_pl + 1] + 1
                end
            end
        end
    end
    H_pl_l,j = H_pl_l,j / Σ_{e=0}^{58} H_pl_l,j[e]
end
```

Line 11: $g_{p\_pl} = Data(i - R_{\mathrm{X}} sin(\frac{2\pi p}{P_{pl}}); j; k - R_{\mathrm{T}} cos(\frac{2\pi p}{P_{pl}}))$

Line 14: $g_{p}\_pl = Data(i; j - R_{\mathrm{Y}} cos(\frac{2\pi p}{P_{pl}}); k - R_{\mathrm{T}} sin(\frac{2\pi p}{P_{pl}}))$

Line 27: $$H_{pl\_l,j} = \frac{H_{pl\_l,j}}{\sum_{e=0}^{58} H_{pl\_l,j}[e]}$$

each g_c by traversing its P_{pl} neighbor pixels from left to right and from top to bottom. The codes obtained, noted $LBP\_g_c\_pl$, are presented as a histogram, noted $H_{pl\_l,j}$ for the plane pl, and for the sub-region j of level l. Algorithm 1 summarizes the mathematical steps to obtain the $H_{pl\_l,j}$ histogram. Eventually, we concatenate horizontally the $H_{pl\_l,j}$ histogram of the XT plane ($H_{\mathrm{XT}\_l,j}$) and the $H_{pl\_l,j}$ histogram of the YT plane ($H_{\mathrm{YT}\_l,j}$) in order to obtain TLBP^{u2} histogram, $H_{l,j}$, sized $59 + 59 = 188$ bins as follows: $H_{l,j} = [H_{\mathrm{XT}\_l,j}\ H_{\mathrm{YT}\_l,j}]$.
3. We further concatenate horizontally the H_{lj} histograms (N TLBP^{u2} histograms for the N sub-region) of the levels considered so as to produce the final PTLBP^{u2} histogram representing our feature vector sized $59 \times 2 \times N$ (Fig. 2c).

3.3 Dimensionality reduction

Dimensionality reduction is effective in removing irrelevant or redundant features, improving result comprehensibility, lowering computational complexity, building better generalizable classifiers and decreasing required storage. In our approach, we apply double dimensionality reduction. The first one is based on feature selection techniques for determining the most discriminating sub-regions. The second one is based on feature extraction techniques for mapping the original feature space of the sub-regions selected to a new feature space with lower dimensions than the ones of the existing space.

3.3.1 Sub-regions selection

To select the sub-regions having the most discriminating features for facial-expression recognition, we can use the Filter and/or Wrapper methods [2, 48]. Compared to Filter methods, Wrapper methods generally obtain better predictive accuracy [27].

Among the most known Wrapper methods,there is Sequential Forward Selection method (SFS) [57]. It starts with the empty set, and sequentially adds the most discriminating features in each iteration. It accesses all possible combinations to choose the best one that maximizes recognition-rate accuracy. Running all combinations is really expensive. So, in our work, we suggest an adapted Wrapper method based on SFS in which we select the most discriminating sub-regions through the pyramidal levels considered.

The process of this method can be formulated by main four steps:

1. Given a subset of N sub-regions of all levels considered $S_{r-1}\{C_1, C_2, ..., C_N\}$, we construct N classifiers corresponding to N sub-regions. For each classifier, we have applied TLBP^{u2} feature detection method on one sub-region C_i ($i = 1, 2, ..., N$) so as to obtain $59 \times 2 = 118$ features (Algorithm 1), Sequential Minimal Optimization (SMO) algorithm [22, 36] with polynomial kernel to classify the expressions, and the subject-independent 10-cross validation to estimate the recognition rate, τ. In fact, to access the classifiers, we decompose the set of video sequences into 10 folds: each fold consists of 90% of subjects for the learning and the remaining 10% of subjects for the test. The recognition rate, τ, is the average of the recognition rates estimated for each fold.
2. We organize the N sub-regions according to their recognition rates in descending order, which is presented by an another subset, noted S_{sr}.
3. We define $N - 1$ recursive combinations of N tidied up sub-regions. As a matter of fact, the first combination corresponds to the first sub-region, and to the second one.

The second combination corresponds to the first combination, and to the third sub-region. Thus, $N-1$ classifiers are generated by TLBP^{u2} features of combination j ($j = [1, 2, \ldots, N-1]$) and SMO algorithm, and the subject-independent 10-cross validation is opted to calculate the recognition rate.

4. We consider only the sub-regions of the best combination that maximizes the recognition rate with minimum of sub-regions. The final result of the discriminating sub-regions is presented in the subset, noted S_r.

The formal description of the method suggested to select the discriminating sub-regions is described by Algorithm 2.

Algorithm 2 The discriminating sub-regions selection

Input : N, the number of total sub-regions.
$S_{r-1}\{C_1, C_2, ..., C_N\}$, the subset of all sub-regions.
Output: S_r, the subset of sub-regions selected.

begin
  $S_{sr} = \{\emptyset\}$ /* the subset of descending order of the sub-regions */
  $S_r = \{\emptyset\}$
  for $i = 1\ to\ N$ **do**
    /* selection of the sub-region of S_{r-1} having the maximum recognition rate τ */
    $C_i = argmax[\tau(S_{r-1})]$
    $S_{sr} = S_{sr} \cup C_i$; $S_{r-1} = S_{r-1} \backslash C_i$
  end
  $S_r = S_r \cup S_{sr}\{C_1\} \cup S_{sr}\{C_2\}$
  $imax = 1$
  $max = \tau(S_r)$ /* max corresponding to the recognition rate of the classifier generated on the first sub-regions combination */
  for $i = 3\ to\ N$ **do**
    $S_r = S_r \cup S_{sr}\{C_i\}$; $\tau_i = \tau(S_r)$
    if $\tau_i > max$ **then**
      $max = \tau_i$; $imax = i$
    end
  end
  for $i = imax + 1\ to\ N$ **do**
    $S_r = S_r \backslash S_{sr}\{C_i\}$
  end
end

3.3.2 Feature extraction

Even after selecting the most discriminating sub-regions, we can find redundant and useless features for facial-expression recognition. Principal Component Analysis (PCA) is the most widely used feature extraction technique for face recognition and facial-expression recognition [5, 18, 58]. In this paper, we use PCA in order to reduce the dimension of $59 \times 2 \times N$ of PTLBP^{u2} features vector (with N is the number of the discriminating sub-region selected i.e., the size of the S_r subset) by projecting it onto d-dimensional space of features where $d << 59 \times 2 \times N$. We start with decomposing the dataset into the learning subset V_l and the test subset V_t presented in terms of $m_l \times n$ matrix and $m_t \times n$ matrix respectively, with m_l

corresponding to the number of learning video sequences, m_t corresponding to the number of test video sequences, and n corresponding to $59 \times 2 \times N$ PTLBP^{u2} features.

For the learning set, we wish to linearly transform the V_l matrix to another matrix of $n_l \times d$ dimension, with $d << n$, PCA[PTLBP^{u2}]. Main four steps can be distinguished:

1. We standardize the V_l matrix by calculating the zero mean of the V_l, noted X_l, as in (1):

$$X_l = V_l - \left(\begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}_{(m_l,1)} \times moy(V_l) \right)_{(m_l,n)} \tag{1}$$

   Where $moy(V_l) = [\frac{1}{m_l}\sum_{i=1}^{m_l} x_{i1} \quad \frac{1}{m_l}\sum_{i=1}^{m_l} x_{i2} \quad \dots \quad \frac{1}{m_l}\sum_{i=1}^{m_l} x_{in}]$;
2. We calculate the matrix of the n eigenvectors (i.e., principal components), noted $E[E_1, E_2, ..., E_n]$. Each eigenvector E_i $(i = 1, 2, ..., n)$ is determined from the eigenvalue, λ_i, maximizing the co-variance matrix of X_l, noted Y and defined by (2) λ_i and E_i is computed through resolving the Eq-System (3):

$$Y_{(n,n)} = cov(X_l) = \frac{1}{m_l} X_l' X_l \tag{2}$$

$$\begin{cases} Y - \lambda_i I_{n,n} = 0 \\ (Y - \lambda_i I_{n,n}) E_i = 0 \end{cases} \tag{3}$$

3. We sort the eigenvalues in descending order, and choose the E_j eigenvectors (with $j = 1, 2, ..., d$) that corresponding to the d largest eigenvalues. Several techniques have been proposed in order to estimate the target dimensionality [14, 23, 31]. In our work, we determine the number d of eigenvectors, that depend on X_l input matrix, using the Maximum Likelihood Estimator [31].
4. We obtain our newly low-dimensional feature space, PCA[PTLBP^{u2}], via projecting X_l onto the matrix of the d sorted eigenvectors, as shown in (4):

$$\text{PCA[PTLBP}^{u2}]_{(m_l,d)} = X_l \times [E_1, E_2, ..., E_d] \tag{4}$$

For the test set, we even calculate the standardizing matrix of V_t, noted X_t, as follow in (5):

$$X_t = V_t - \left(\begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}_{(m_t,1)} \times moy(V_l) \right)_{(m_l,n)} \tag{5}$$

Then, we project X_t onto the d eigenvectors detected as in step 4 so as to obtain the test features.

3.4 Expression classification

The performance of our approach has been evaluated, using two different supervised learning techniques: SVM [52] and C4.5 decision trees [38].

SVM developed by Vapnik [52] is inherently a binary classifier, and can also be used for multi-classification problems. It has a high generalization performance regardless of prior knowledge, even when the dimension of the input space is very large. SVM consists of finding the maximal margin separating hyperplane with respect to a training set that correctly classifies the data points, using a specific $K(F_i, F)$ kernel function. In our work, we use two kernels: the polynomial kernel defined as $(F_i' \times F + 1)^d$, and the Radial Basis Function kernel (RBF) defined as $\exp(-\gamma \left\| F_i' - F \right\|^2)$. F_i is the feature vector of i^{th} video sequence in the training set, F is the feature vector of video sequence in the test set to be classified, $d > 0$ is the degree of the kernel, and $\gamma > 0$ is the free parameter that allows changing the size of the RBF kernel.

A standard SVM seeks to find a margin separating all positive and negative examples. However, this can lead to poorly fit models if some examples are mislabeled or extremely unusual. Therefore, we apply the "Soft margin" technique [6] that minimizes the number of errors by loosening constraints on training vectors. This loosening requires the definition of a constant C to control the trade-off between the number of misclassifications and the margin width. Furthermore, we apply Sequential Minimal Optimization (SMO) algorithm [22, 36] so as to solve the quadratic programming (QP) optimization problem that arises during the SVM training. Indeed, SMO splits this problem into a series of smaller possible sub-problems which are later solved analytically.

We use two strategies for the multi-class classification problem: (i) "1-vs-1" that decomposes the problem of f classes into $\frac{f\cdot(f-1)}{2}$ binary classifiers, and (ii) "1-vs-all" in which one binary SVM is built for each class to separate members of that class from members of other classes.

C4.5 developed by Ross Quinlan [38] is an algorithm, allowing generating a decision tree (from which decision rules of type "if ... else" are derived) through the "gain ratio" as a splitting criterion, and through post-purring technique to reduce the size of a decision tree and avoid over-fitting. Formally, let c denotes the number of classes; $S_{i-1} = \{s_{(i-1)1}, s_{(i-1)2}, ..., s_{(i-1)\beta}\}$ is the partition that we seek to split its β nodes, noted $s_{(i-1)k}$ (with $k = 1..\beta$); $S_i = \{s_{i1}, ..., s_{i\alpha}\}$ is the partition obtained through splitting the node $s_{(i-1)k}$ (with $k = 1..\beta$); and $p(S, j)$ is the proportion of instances in the partition S ($S = S_{i-1} or S_i$) which are assigned to the $j - th$ class. Therefore, we calculate the Shannon's entropy, noted $I(S)$, as follows in (6):

$$I(S) = -\sum_{j=1}^{c} p_j \log_2(p_j) \tag{6}$$

Accordingly, given $n_{s_{ik}}$ as the number of video sequences in node k of the partition S_i, and $n_{S_{i-1}}$ as the number of video sequences in the partition S_{i-1}, the "gain ratio" measurement of S_i, noted $\Im(S_i)$, is calculated as in (7):

$$\Im(S_i) = \frac{I(S_{i-1}) - I(S_i)}{-\sum_{k=1}^{\alpha} \left(\frac{n_{s_{ik}}}{n_{S_{i-1}}}\right) \log_2 \left(\frac{n_{s_{ik}}}{n_{S_{i-1}}}\right)} \tag{7}$$

All nodes of each partition have been splitted according to each feature. We consider the best feature selected as the one that gives the best separation of video sequences in accordance with the classes, showing the best "gain ratio" measurement. Then, we repeat recursively the splitting process on each node until we do not have features available for partitioning, and/or until the node is "pure" or "almost pure" i.e., all video sequences in the same node belong to a single class. Finally, we apply the post-purring technique, known as a sub-tree replacement, whose aim is to reduce the number of nodes.

4 Experiments

To validate the approach proposed, we reserve the next sections for presenting firstly the choice of facial-expression databases, secondly the defined experimental conditions, and finally the different conducted experiments.

4.1 Facial-expression databases

In this work, we consider the six universal facial expressions (joy, surprise, disgust, sadness, anger and fear) [11]. We use two different facial-expression databases: the extended Cohn-Kanade (CK+) database [21] and the MMI database [35].

CK+ database [21] is the extended version of CK database. It consists of 593 video sequences of 123 subjects aged from 18 to 30 years old of African-American, Asian and Latino origins. Only 327 sequences are annotated: 309 by the six universal facial expressions and 18 by the contempt expression. All of the sequences contain one frontal face, representing a posed facial expression. The video sequences consist of two moments: Onset and Apex (OA). In the Onset moment, the facial face changes from neutral state to its maximum expressive state. In the Apex moment, the facial face stagnates at its maximum expressive state. The frames of sequences have been digitized into either 640×490 or 640×480 pixels with a 8-bit gray scale value or a 24-bit color value as in Fig. 3a. We have tried to annotate the 266 non-annotated video sequences of CK+ database based on the decision taken by a number of 13 participants (men and women) of different social levels. Each participant follows the video sequence, and attributes their decision that corresponds to one of the six universal facial expressions or the neutrality expression in case of confusion. Indeed, for each video sequence, seven occurrences of expressions, noted occ_i with $i = 1, 2, ..., 7$ are calculated. We consider only the sequences firstly having the difference between the maximal occurrence occ_j of expression j and each occurrence occ_i ($i \in \{1, 2, ..., 7\}$ with $i \neq j$) that must exceed the majority number of the subjects taking part in the annotation process, and secondly the maximal occurrence does not correspond to the neutrality expression. We use 161 video sequences personally annotated from 266 video sequences non-annotated plus 309 video sequences already annotated by the universal expressions. We have finally obtained a total of 470 well-annotated video sequences.

MMI database [35] contains both static images and video sequences taken from posed expressions, including possible head movements. It consists of over 2900 videos and high-resolution still images of 88 subjects of both sexes (34% female), aged from 19 to 62 years old, with either a European, Asian, or South American ethnicity. Only 204 from 2392 sequences are annotated by the six universal facial expressions. The facial face of each video sequence has gone through with three moments (OAO): from neutral to expressive state i.e., the Onset moment; keeping in the expressive state i.e., the Apex moment; and from expressive to neutral state i.e., the Offset moment. MMI database includes more subject variation than the CK+ database i.e., faces may be partially occluded by beard, mustache and glasses. The frames of sequences have been digitized by a 24-bit color value (frontal, profile and dual-view recordings) of 720×576 pixels as in Fig. 3b. For our experiments, six video sequences have been rejected following an inaccurate detection of the face by applying the Viola and Jones algorithm. We have used only 198 video sequences in which we have manually selected the Onset and the Apex moments so that the video sequence will start with a neutrality face and end with an expressive face.

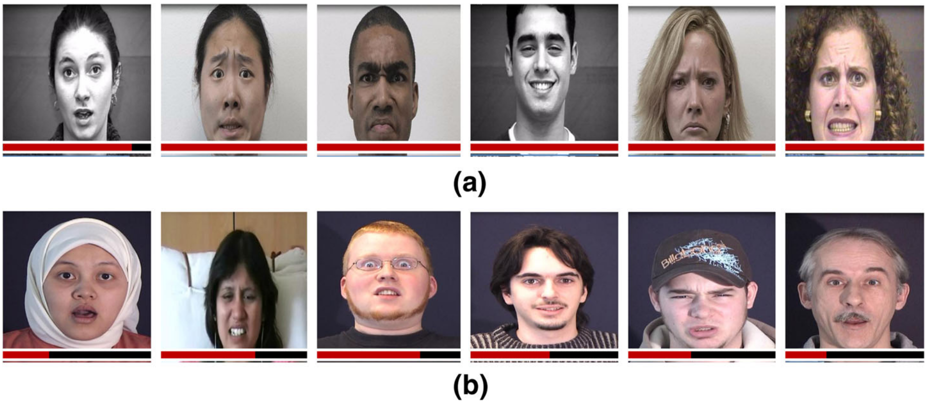

Fig. 3 Examples of frames from CK+ and MMI databases

4.2 Experimental conditions

We distinguish two series of experiments. The first one considers the experiments performed during the off-line stage in that we determine the levels of pyramid used for detecting PTLBP^{u2} features, and select the discriminating sub-regions by training the suggested method in Section 3.3.1. The second series presents the experiments performed during the on-line stage in which we present two series of experiments: (i) assessment of facial-expression recognition approach under a lab-controlled environment, and (ii) evaluation of facial-expression recognition approach under an uncontrolled environment. In the first series of the on-line experiments, we use CK+ and MMI databases separately, and we apply the subject-independent 10-cross validation to access the classifiers. Precisely, we define 10 folds, and for each fold, we select 90% of subjects for the learning, and the remaining 10% subjects for the test i.e., the faces used in the learning phase do not contribute to the test phase. In the second series of the on-line experiments, we use cross-databases. Indeed, we use CK+ as the training set, and MMI as the test set, and vice versa. As validation indicators, we use the recognition rate and the Area Under Curve (AUC) that is calculated referring to ROC curve [1] so as to evaluate the performance of all experiments. To compute AUC, we plot only one ROC graph for each expression each considered as the positive class and all other classes as the negative one.

We only present the results performed in the test phase for all off-line and on-line experiments. They allow us to evaluate the effectiveness and robustness of the approach proposed. The program of face detection and tracking, feature detection, and dimensionality reduction have been developed, using Matlab R2015a. Weka is applied to generate and evaluate the prediction models used in expression classification. A PC with Intel (R) Core (TM) i3 CPU 2013 GHz and 3 GB of Random Access Memory (RAM) has been used to perform the experiments.

4.3 Off-line experiments

We have carried out off-line experiments to evaluate the performance of the approach proposed. These experiments have been conducted using CK+ database with 309 video sequences. First, we seek to determine the pyramidal representation i.e., the best combination of levels that can show the most discriminating features. We specify only three levels: level 1, level 2 and level 3. Hence, Four combinations are distinguished: "level 1 + level 2", "level 1 + level 3", "level 2 + level 3", and "level 1 + level 2 + level 3". For each combination, three classifiers have been generated. They are SVM with polynomial kernel, SVM with RBF kernel, and C4.5, using all sub-regions of each level. Moreover, we use the two multi-class classification strategies which are "1-vs-1" and "1-vs-all". Several parameters of SVM can affect the efficiency of our classifiers. Among these parameters, we identify the degree d and C of the polynomial kernel, and the parameters γ and C of the RBF kernel. These parameters is determined by an empirical study. Using the subject-independent 10-cross validation, we determine the best combination of $d \in [1, 4]$ and $C \in [0, 8]$ for the polynomial kernel, and of $C \in [0, 8]$ and $\gamma \in [0.001, 0.015]$ for the RBF kernel. The results obtained using the best parameters for each classifier are summarized in Table 1.

We note that "Num.Reg", "Num.Feat", "Best.Par (Poly)", "Best.Par (RBF)", "Best.Str" and "Comp. time" refer to the number of sub-regions used in detecting PTLBP^{u2} features, the number of features, the best parameters of SVM with polynomial kernel, the best parameters of SVM with RBF kernel, the best strategy of multi-class classification and the computation time of feature detection respectively.

As shown in Table 1, we have observed that all classifiers generated by SVM with polynomial kernel and "1-vs-1" strategy always record the recognition rates (reaching 90.6%) remarkably higher than those generated by SVM with RBF kernel (reaching 88.6%), and C4.5 (reaching 71.5%). For that, we have considered only the results produced by SVM with polynomial kernel for determining the best combination of levels, and therefore for the remaining studies conducted in off-line experiments. More precisely, classifier 3 presents the highest recognition rate which reaches 90.6%, and the best computation time of features that achieves 1.8 seconds. For classifiers 1 and 2, the computation time of features is higher than that obtained by classifier 3 although the number of sub-regions of the "level 1 + level 2" and "level 1 + level 3" is lower than that of the "level 2 + level 3". This decrease of the computation time of features can be explained by the variability of the number of center pixels calculated referring to sub-regions selected with different sizes. More precisely, for a video sequence with k frames, r sub-regions (each sized $h \times w$ pixels) and $R_X = R_Y = R_T = 1$, the first center pixel has $(R_X + 1, R_Y + 1, R_T + 1)$ coordinates, and the last center pixel has $(h - R_X, w - R_Y, k - R_T)$ coordinates i.e., the total number of center pixels is equal to $(h - 2) \times (w - 2) \times (k - 2) \times r$. For example, given a video sequence with 11 frames, for level 1 (4 sub-regions, each among them sized 32×32 pixels), the total number of center pixels calculated is equal to $(32 - 2) \times (32 - 2) \times (11 - 2) \times 4 = 32400$. Likewise, for level 2 (16 sub-regions, each among them sized 16×16 pixels), it is equal to 28224, and for level 3 (64 sub-regions, each among them sized 8×8 pixels), it is equal to 20736. The total number of center pixels for levels 2 and 3 is much lower than those for level 1. Hence, we choose to use the sub-regions of level 2 and level 3 to calculate the features in the subsequent experiments.

Using all sub-regions of level 2 and level 3 is costly in terms of spatial and temporal complexity. For that, we propose an algorithm based on the Wrapper method to select the sub-regions having the best discriminating features for facial-expression recognition. The learning is performed with SVM (polynomial kernel with $d = 1$, $C = 1$ and "1-vs-1" strategy) on the sub-regions selected, and the performance is estimated through the subject-independent 10-cross validation. We calculate 79 combinations of sub-regions from which we extract the one having the highest facial recognition rate and AUC. We have found out that the combination performed by 8 sub-regions of level 2 and 25 sub-regions of level 3

Table 1 Evaluation results of different combinations of the pyramid levels

N°	Combinations of levels	Num. Reg	Num. Feat	Best.Par (Poly) (C, d)	Best.Par (RBF) (C, γ)	Best.Str	Recognition rate (%)			Comp. time (s)
							SVM Poly	SVM RBF	C4.5	
1	level 1+level 2	20	2360	(1,1)	(5,0.003)	1-vs-1	88.0	88.6	71.5	2.5
2	level 1+level 3	68	8024	(1,1)	(6,0.001)	1-vs-1	89.3	88.0	63.8	2.2
3	level 2+level 3	80	9440	(1,1)	(5,0.001)	1-vs-1	90.6	88.3	64.7	1.8
4	level 1+level 2+ level 3	84	9912	(1,1)	(3,0.001)	1-vs-1	89.6	88.0	65.4	2.1

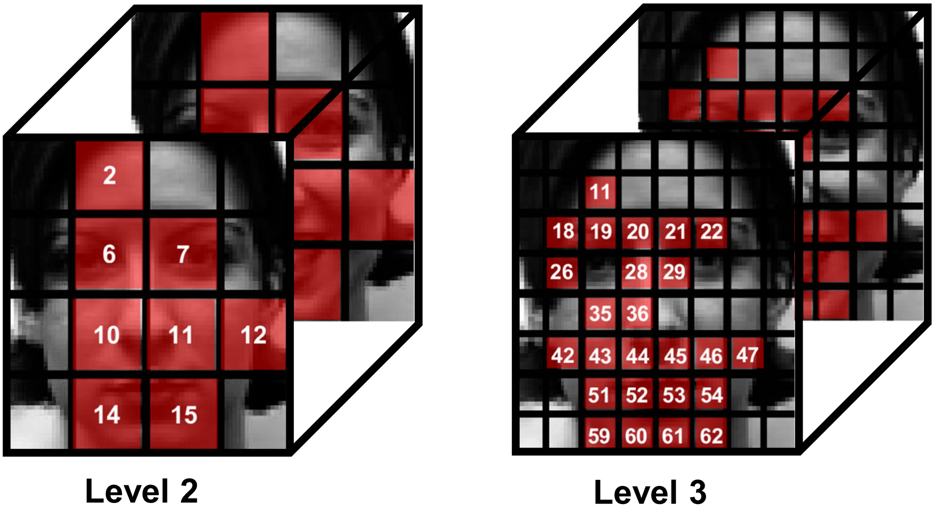

Fig. 4 Example of sub-regions selected of level 2 and level 3, taken from CK+ database

records the highest recognition rate (95.14%). An example of the result of sub-regions selected- that is produced by Algorithm 2- is shown in Fig. 4. Like any other method of features selection, the method suggested can neglect some discriminating sub-regions. However, it differs from the other techniques via selecting sub-regions from two different levels, possibly leading to have a double selection in the same sub-regions, or a part of the same sub-regions. Our method has a double role. It firstly selects the most important sub-regions, and secondly weighed the most discriminating sub-regions in this selection. As shown in Fig. 4, for instance, the sub-region 10 of level 2 is re-selected in level 3 (sub-regions 35, 36, 43 and 44). The 33 sub-regions selected (8 of level 2 and 25 of level 3) are mainly around the mouth, cheeks, eyes and eyebrows.

4.4 On-line experiments

Based on the result found on the off-line experiments, we carry out our on-line experiments to evaluate the performance of the approach proposed. It will be interesting to evaluate all classifiers generated by SVM or C4.5 before and after selecting the discriminating sub-regions, and before and after extracting features via PCA. The result of the classifiers generated by C4.5 before applying PCA is not efficient. Thus, in the subsequent experiments, we present only the results of four classifiers. In the first one, we have used all sub-regions of level 2 and level 3. In the second one, we have considered only the 33 discriminating sub-regions selected. We use PTLBP^{u2} for detecting features, and SVM for expression classification. In the third one, we have applied PCA on the feature vector generated by PTLBP^{u2} on the 33 discriminating sub-regions selected and SVM. The fourth one is also produced by PCA on PTLBP^{u2}- using the 33 discriminating sub-regions selected- and C4.5.

4.4.1 Facial-expression recognition under a lab-controlled environment

In this section, we evaluate our different classifiers, using CK+ and MMI databases separately. All experiments have been expressed in terms of the subject-independent 10-cross

Table 2 Experimental results of the approach proposed, using the "v1_CK+" database

N°	Features vector	Number of sub-regions	Size of feature vector	Learning technique	Recognition rate %	Area Under Curve (AUC)
1	PTLBP^{u2}	All	9940	SVM	90.6	0.96
2	PTLBP^{u2}	33	3894	SVM	95.1	0.98
3	PCA [PTLBP^{u2}]	33	45	SVM	97.1	0.99
4	PCA [PTLBP^{u2}]	33	45	C4.5	98.1	0.99

validation, and the performance is measured by the recognition rate and AUC. In particular, for classifiers 3 and 4, we apply Maximum Likelihood Estimator (MLE) [31] to find the number d of principal components. This number is variable, and depends on the learning subset, V_l (Section 3.3.2). Therefore, we obtain 10 values of d corresponding to the 10-folds in which we consider the mean of all values found.

Evaluation of different classifiers using CK+ database For evaluating our approach, we use two versions of CK+ database: the first one, called "v1_CK+", contains only the 309 already annotated video sequences, and the second one, called "v2_CK+", contains the 309 video sequences used in "v1_CK+", and 161 other video sequences annotated in the present work. Tables 2 and 3 represent the evaluation results of the approach proposed, using the "v1_CK+" and the "v2_CK+" databases respectively. For "v1_CK+", comparing classifier 1 (before selecting sub-regions) and classifier 2 (after selecting sub-regions), we find that using the 33 sub-regions selected is better than using all sub-regions in which the recognition rate has improved from 90.6 to 95.1%. Applying PCA on PTLBP^{u2} feature vector, using the 33 discriminating sub-regions (PCA [PTLBP^{u2}]) increases the recognition rate regardless of the learning technique used. Indeed, classifier 4 records the highest recognition rate of 98.1%. This improvement is due to the size of the new space that consists of a very small number of features compared to that of the PTLBP^{u2}, using only the 33 sub-regions selected. Similarly, for "v2_CK+", classifier 4 illustrates the highest recognition rate of 98.9%, even higher than the results found by classifier 4 generated by using CK+ with 309 video sequences. However, the performance of classifier 2 (generated by PTLBP^{u2}, using the 33 sub-regions selected, and SVM learning technique) via "v2_CK+" database falls remarkably compared to classifier 2 performed by using "v1_CK+" database. On the other hand, the values of AUC (0.95-0.99) found by using 309 or 470 video sequences show the specificity of all classifiers, and prove the significance of the PTLBP^{u2} features.

Evaluation of different classifiers using MMI database We evaluate the approach proposed via using two versions of MMI database composed of 198 video sequences: the

Table 3 Experimental results of the approach proposed, using the "v2_CK+" database

N°	Features vector	Number of sub-regions	Size of feature vector	Learning technique	Recognition rate %	Area Under Curve (AUC)
1	PTLBP^{u2}	All	9940	SVM	87.2	0.95
2	PTLBP^{u2}	33	3894	SVM	91.7	0.97
3	PCA [PTLBP^{u2}]	33	50	SVM	97.2	0.99
4	PCA [PTLBP^{u2}]	33	50	C4.5	98.9	0.99

Table 4 Experimental results of the approach proposed, using the "v1_MMI" database

N°	Features vector	Number of sub-regions	Size of feature vector	Learning technique	Recognition rate %	Area Under Curve (AUC)
1	PTLBP^{u2}	All	9940	SVM	59.6	0.82
2	PTLBP^{u2}	33	3894	SVM	63.6	0.86
3	PCA [PTLBP^{u2}]	33	33	SVM	80.1	0.96
4	PCA [PTLBP^{u2}]	33	33	C4.5	97.5	0.99

first one, named "v1_MMI", considers only the Onset and Apex moments (OA) that are selected manually, and the second one, named "v2_MMI", consists of the three Onset, Apex and Offset moments (OAO). Tables 4 and 5 represent the results obtained by our classifiers, using "v1_MMI" and "v2_MMI" databases respectively. Using the CK+ database, classifier 2 (after selecting sub-regions) is better than classifier 1 (before selecting sub-regions) but using the MMI database, the performance of classifier 2 falls remarkably from 95.1–91.7% to 63.6–58.6% of recognition rates. This decrease in facial-expression recognition is explained by the fact that the MMI database presents higher variabilities than the CK+ database. Despite the irrelevance of recognition rate, the AUC results prove the performance of classifier 2. Nevertheless, the performance of classifiers 3 and 4 (generated by PCA on TPLBP^{u2}, using the 33 sub-regions selected) improves, especially when we have applied C4.5 as learning technique (97–97.5% of recognition rate and 0.98–0.99 of AUC). The results obtained by classifier 4, using MMI database almost coincide with those found by classifier 4, using CK+ database which demonstrates the importance and the flexibility of the approach proposed. Hence, using PCA on the TPLBP^{u2} feature space improves the accuracy of facial-expression recognition.

Positioning of the approach proposed relative to related works We conduct a comparison of our approach with other state-of-the-art approaches in terms of the size of feature vector (Size.Vector), the number of video sequences (Seq.Numb), the number of class expressions (Class.Numb) and the performance measure (Perf.Meas), providing the recognition rate (Recog.Rate) using the CK+ and MMI databases. Given that the experimental conditions are not the same, it is difficult to make a quantitative comparison between our approach and the state-of-the-art facial-expression recognition. Most of literature approaches apply n-fold cross-validation (2-folds, 4-folds, 10-folds) as an evaluation indicator. Generally, using 10-folds gives the chance of having more effective results than using 2-folds or 4-folds. Thus, we find it interesting to evaluate our approach through

Table 5 Experimental results of the approach proposed, using the "v2_MMI" database

N°	Features vector	Number of sub-regions	Size of feature vector	Learning technique	Recognition rate %	Area Under Curve (AUC)
1	PTLBP^{u2}	All	9940	SVM	52.5	0.80
2	PTLBP^{u2}	33	3894	SVM	58.6	0.83
3	PCA [PTLBP^{u2}]	33	29	SVM	76.1	0.95
4	PCA [PTLBP^{u2}]	33	29	C4.5	97.0	0.98

Table 6 Positioning of the approach proposed relative to the literature ones, using CK+ and MMI databases

	Size. Vector	Seq. Numb CK+/MMI	Class. Numb	Dyn	Perf. Meas	Recog. rate (%)
Pu et al. [37]	–	327^+ / –	7 / –	No	9-folds	96.4 / –
Khan et al. [26]	590	309^+ / 392^+	6 / 3	No	10-folds	96.7 / 91.4
Sanchez et al. [41]	1200	348^+ / 96^+	6 / –	Yes	4-folds	92.8 / 82.7
Wang et al. [55]	–	327^+ / 196^*	7 / 6	Yes	10-folds	81.7 / 52.1
Zhao et al. [62]	12744	374^+ / –	6 / –	Yes	2-folds	94.4 / –
Zhao et al. [63]	–	374^+ / –	6 / –	Yes	2-folds	93.8 / –
Wang et al. [56]	531	309^+ / –	6 / –	Yes	10-folds	86.6 / –
Ji et al. [20]	2752	348^+ / 199^+	6 / –	Yes	10-folds	93.7 / 95.0
Fan et al. [12]	1860	– / 203^+	– / 6	Yes	10-folds	– / 74.3
Chen et al. [3]	2382	327^+ / –	7 / –	Yes	Leave-one-subject-out	89.6 / –
Ours	45 / 33	309^+ / 198^+	6 / 6	Yes	10-folds	**98.1 / 97.5**
Ours	50 / 29	470^+ / 198^*	6 / 6	Yes	10-folds	**98.9 / 97.0**
Ours	45 / 33	309^+ / 198^+	6 / 6	Yes	2-folds	**96.8 / 92.4**
Ours	50 / 29	470^+ / 198^*	6 / 6	Yes	2-folds	**96.9 / 94.9**

$^+$: Onset-Apex (OA) *: Onset-Apex-Offset (OAO)

the subject-independent 2-cross validation as well as the subject-independent 10-cross validation to be positioned than the literature. The results are presented in Table 6.

As the most related works, the performance using CK+ database is more important than that of using MMI database. Applying Onset-Apex-Offset (OAO) video sequences makes facial-expression recognition confusing. In fact, the recognition rate is always less than that of using Onset-Apex (OA) video sequences. For instance, Wang et al. [55] have used Onset-Apex-Offset video sequences to evaluate their approach showing about 52% of recognition rate unlike the approaches proposed by Sanchez et al. [41] and Ji et al. [20], using the Onset-Apex video sequences, that go up to 82.7% and 95.0% of recognition rate respectively. The dimensionality of the feature space proposed is very low. Compared to Zhao et al. [62] that use 12744 features, and Wang et al. [56] that use 531 features, our approach is based only on 23–51 features, and have produced the best performance. Although the experimental conditions are not the same, we can clearly conclude that our approach, using either the subject-independent 10-cross validation or the subject-independent 2-cross validation, is considered as a strong competitor not only to the dynamic approaches but also to the static ones in the literature.

4.4.2 Facial-expression recognition under an uncontrolled environment

Encouraged by the results of the evaluation of different classifiers under a lab-controlled environment, we devote this section to evaluate the best classifier obtained- that is generated by PCA on PTLBP^{u2}, using the 33 discriminating sub-regions selected, and C4.5 learning technique ("PCA [PTLBP^{u2}] / C4.5")- under an uncontrolled environment. The classifier must be able to classify facial expressions of a subject who does not belong to the same capture environment. This experiment simulates the real life situation when the system is

Table 7 Evaluation results of "PCA [PTLBP^{u2}] / C4.5" classifier under an uncontrolled environment

Scenario number	Learning/test database	Seq. Numb learning/test	Size of feature vector	Recognition rate (%)	Area under curve
1	"v1_CK+"/"v1_MMI"	309/198	45	92.9	0.95
2	"v1_CK+"/"v2_MMI"	309/198	45	92.4	0.95
3	"v1_MMI"/"v1_CK+"	198/309	35	97.7	0.99
4	"v2_MMI"/"v1_CK+"	198/309	35	93.8	0.96

used to recognize facial expressions on the unseen data (different resolutions, color depth, ethnicity, gender, age). It is performed in four different scenarios: the first one considers "v1_CK+" as a learning database, and "v1_MMI" as a test database; the second one considers "v1_CK+" as a learning database, and "v2_MMI" as a test database; the third one considers "v1_MMI" as a learning database, and "v1_CK+" as a test database; and the last one considers "v2_MMI" as a learning database, and "v1_CK+" as a test database. The recognition rate and AUC are used to validate these scenarios of experiments. The results obtained are presented in Table 7.

The performance of classifiers of the four scenarios is efficient since it shows between 92.4% and 97.7% of recognition rate, and between 0.95 and 0.99 of AUC. The classifier of scenario 3, that has used "v1_MMI" as a learning database, and "v1_CK+" as a test database, records the best result. Figure 5 shows the recognition rate of each facial expression obtained by each scenario. The two classifiers of scenario 1 and scenario 2 having CK+ as a learning database show that the expressions of joy, surprise, fear and sadness record the highest recognition rate of up to 100% in contrast with the remaining expressions, resulting in low recognition rates ranging between 72.7% and 83.6%. The expression of surprise is misclassified when using MMI as a learning database (86.4 and 63.1% of recognition rate respectively for scenario 3 and scenario 4). It is also observed that the expressions of joy, fear and sadness are always recognized independently of the choice of learning and test database. For scenarios 1 and 2, the misclassification of disgust and anger expressions can be explained by the striking similarity of these expressions in CK+ database. In fact, in the annotation process of the 266 video sequences of CK+ database, the majority of participants have been confused about the choice between disgust and anger expressions. For

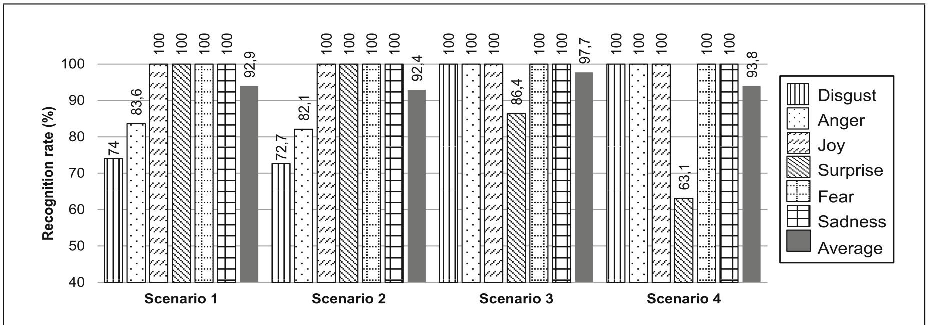

Fig. 5 The performance obtained by each facial expression of "PCA [PTLBP^{u2}] / C4.5" classifier on scenarios 1, 2, 3 and 4

Table 8 Positioning of the approach proposed relative to the literature ones, using CK+ as the learning database, and MMI as the test database

	Features	Type of data	Recognition rate %
Shan et al. [43]	LBP	Image	51.1
Mayer et al. [33]	Candide-III	Image	60.3
Zhang et al. [61]	LBPH+HOG	Image	66.9
Fan et al. [12]	PHOG-TOP+dense optical flow	Video	58.7
Guo et al. [17]	Sparse representation	Video	91.9
Ours	PCA [PTLBP^{u2}]	Video	92.9

scenarios 3 and 4, the misclassification of surprise expression is due to the number of surprise video sequences in MMI database (38 video sequences) which is very low compared to that in CK+ database (83 video sequences). Therefore, these series of experiments proves the comprehensiveness of the approach proposed under an uncontrolled environment.

Scenario 1 is used for positioning our approach, referring to other related works that have considered cross-databases i.e., the facial-expression recognition under an uncontrolled environment, as shown in Table 8. Relative to the other works served on the video sequences [12, 17], and even on the static images [33, 43, 61], our approach still records better recognition performance with 92.4%, using CK+ as the learning database, and MMI as the test database. We have further observed that the recognition rate obtained by using cross-databases have been much lower than those using databases separately. For instance, Fan et al. [12] have obtained 58.7% of recognition rate, using CK+ as the learning database, and MMI as the test database. Unlike the recognition rates of evaluation, each database separately has achieved 83.7%. An interesting approach using cross-databases evaluation is presented in [17] with 91.9% of recognition rate where the CK+ database is applied to construct dynamic-expression atlas sequences based on sparse representation groupwise registration. These atlas sequences are then used to guide the facial-expression recognition on the MMI database. The atlas constructed can capture the facial-appearance movements among the population, which reflects the recognition of facial expressions under an uncontrolled environment. However, the result obtained has also decreased from 96.8% by using only CK+ database to 91.9% by considering CK+ as the learning database and MMI as the test database. This decrease is due to the different conditions of capture for each environment i.e., for each database such as the difference in the variability of illumination, the facial shape, the resolution and the head pose.

5 Conclusion and perspectives

We have proposed a Dynamic Feature-Based approach for facial-expression recognition taken from frontal-face video sequences. It consists in defining a temporal low-dimensional feature space through the second and the third levels of a pyramidal representation, using the discriminating sub-regions selected. We have first suggested an adopted Wrapper method to select the most discriminating sub-regions. We have then calculated dynamic features for each sub-region according to the Temporal orthogonal planes (XT and YT) of uniform Local Binary Pattern (TLBP^{u2}). We have further transformed the final feature space obtained (PTLBP^{u2}) into a low-dimensional feature space in accordance with PCA. Three major contributions could be distinguished. Firstly, Our approach proves a reliable recognition rate under a controlled and an uncontrolled environment. Indeed, under a lab-controlled environment, the effectiveness of our approach reaches a 98.9% of recognition rate, using each database separately. Under an uncontrolled environment, the testing of the approach proposed- using CK+ as a learning database, and MMI as a test database, and vice versa-shows promising results that record recognition rates between 92.9 and 97.7%. Secondly, the use of a pyramidal representation on different levels has allowed weighing certain sub-regions rather than other ones. Then, having only used the discriminating sub-regions has decreased the number of features, and provided us with a profit in the memory space required. Finally, the low-dimensional representation, calculated by PCA, has reduced the redundant and useless features, and made them more discriminating for the facial-expression recognition. Thus, under a lab-controlled environment, the result has improved remarkably from 95.1 to 98.1%, using the CK+ database, and from 63.6 to 97.5%, using the MMI database. It has also shown the highest recognition rate (reaching 97.7%) under an uncontrolled environment.

As perspectives, we could test our approach, using a variety of video sequences (displaying spontaneous expressions, micro-expressions, etc.), taking into account a variation of face poses, a total occlusion, etc. Eventually, we could adopt our approach to classify other non-universal expressions like “worry”, “pain” and “boredom”, having more different intensity degrees than those of the universal facial expressions.

Acknowledgements The authors are grateful to Sofiene HADDED, teacher of English at the Faculty of Economics and Management of Sfax, Tunisia for having proofread the manuscript.

References

1. Branco P, Torgo L, Ribeiro RP (2016) A survey of predictive modeling on imbalanced domains. ACM Comput Surv 49(2):1–50. article No. 31
2. Chandrashekar G, Sahin F (2014) A survey on feature selection methods. Comput Electr Eng 40(1): 16–28
3. Chen J, Chen Z, Chi Z, Fu H (2016) Facial expression recognition in video with multiple feature fusion. IEEE Trans Affect Comput PP(99):1–12
4. Cheng H, Shi X (2004) A simple and effective histogram equalization approach to image enhancement. Digit Signal Process 14(2):158–170
5. Chung KC, Kee SC, Kim SR (1999) Face recognition using principal component analysis of gabor filter responses. In: Proceedings of the international workshop on recognition, analysis, and tracking of faces and gestures in real-time systems. IEEE Computer Society, pp 53–57
6. Cortes C, Vapnik V (1995) Support vector networks. J Mach Learn 20(3):237–297
7. De La Torre F, Cohn JF (2011) Facial expression analysis. In: Visual analysis of humans: looking at people. Springer, London, pp 377-409
8. Deng B, Jin LW, Zhen LX, Huang JC, Deng HB (2005) A new facial expression recognition method based on Local Gabor Filter bank and PCA plus LDA. Inf Technol IT 11:86–96
9. Donia MMF, Youssif AAA, Hashad A (2014) Spontaneous facial expression recognition based on histogram of oriented gradients descriptor. Comput Inf Sci 7:31–37
10. Doretto G, Chiuso A, Nian YN, Soatto S (2003) Dynamic textures. Int J Comput Vis 51(2):91–109

11. Ekman P (1972) Universals and cultural differences in facial expressions of emotion. University of Nebraska Press Lincoln
12. Fan X, Tjahjadi T (2015) A spatial-temporal framework based on histogram of gradients and optical flow for facial expression recognition in video sequences. Pattern Recogn 48(11):3407–3416
13. Fu X, Wei W (2008) Centralized binary patterns embedded with image Euclidean distance for facial expression recognition. In: Proceedings of the 2008 fourth international conference on natural computation ICNC '08, vol 04. IEEE Computer Society, Washington, DC, pp 115–119
14. Fukunaga K, Olsen DR (1971) An algorithm for finding intrinsic dimensionality of data. IEEE Trans Comput C-20(2):176–183
15. Gonzalez RC, Woods RE (2008) Digital image processing, 3rd edn. Upper Saddle River, Prentice Hall
16. Gritti T, Shan C, Jeanne V, Braspenning R (2008) Local features based facial expression recognition with face registration errors. In: 2008 8th IEEE International conference on automatic face gesture recognition, pp 1–8
17. Guo Y, Zhao G, Pietikäinen M (2016) Dynamic facial expression recognition with Atlas construction and sparse representation. IEEE Trans Image Process 25(5):1977–1992
18. Happy SL, George A, Routray A (2012) A real time facial expression classification system using local binary patterns. In: 2012 4th International conference on intelligent human computer interaction (IHCI), pp 1–5
19. IMOTIONS - BIOMETRIC RESEARCH PLATFORM (2016) Facial expression analysis: The complete pocket guide. https://imotions.com/blog/facial-expression-analysis
20. Ji Y, Idrissi K (2012) Automatic facial expression recognition based on spatiotemporal descriptors. Pattern Recogn Lett 33(10):1373–1380
21. Kanade T, Cohn JF, Tian Y (2000) Comprehensive database for facial expression analysis. In: Proceedings Fourth IEEE international conference on automatic face and gesture recognition, pp 46–53
22. Keerthi SS, Shevade SK, Bhattacharyya C, Murthy KRK (2001) Improvements to Platt's SMO algorithm for SVM classifier design. Neural Comput 13(3):637–649
23. Kégl B (2002) Intrinsic dimension estimation based on packing numbers. In: Advances in neural information processing systems. Cambridge, pp 833–840
24. Khan RA (2013) Detection of emotions from video in non-controlled environment. University of Claude Bernard of Lyon, Phd thesis
25. Khan RA, Meyer A, Konik H, Bouakaz S (2012) Human vision inspired framework for facial expressions recognition. In: 2012 19th IEEE international conference on image processing, pp 2593–2596
26. Khan RA, Meyer A, Konik H, Bouakaz S (2013) Framework for reliable, real-time facial expression recognition for low resolution images. Pattern Recogn Lett 34(10):1159–1168
27. Kohavi R, John GH (1997) Wrappers for feature subset selection. Artif Intell 97(1):273–324
28. Kotsia I, Zafeiriou S, Pitas I (2008) Texture and shape information fusion for facial expression and facial action unit recognition. Pattern Recog 41(3):833–851. Part Special issue: Feature Generation and Machine Learning for Robust Multimodal Biometrics
29. Kumari J, Rajesh R, Pooja KM (2015) Facial expression recognition: a survey. Procedia Comput Sci 58:486–491. Second international symposium on computer vision and the internet
30. Lee TZ, Bong DBL (2016) Analysis of Bit-Plane images by using Principal Component on face and palmprint database. Pertanika J Sci Technol 24(1):191–203
31. Levina E, Bickel PJ (2004) Maximum likelihood estimation of intrinsic dimension. In: Advances in neural information processing systems. Cambridge, pp 777–784
32. Littlewort G, Bartlett MS, Fasel I, Susskind J, Movellan J (2006) Dynamics of facial expression extracted automatically from video. Image Vis Comput 24(6):615–625. Face processing in video sequences
33. Mayer C, Eggers M, Radig B (2014) Cross-database evaluation for facial expression recognition. Pattern Recogn Image Anal 24(1):124–132
34. Mliki H, Hammami M, Ben-Abdallah H (2013) Mutual information-based facial expression recognition. In: 2013 Sixth international conference on machine vision, society of photo-optical instrumentation engineers (SPIE), vol 9067
35. Pantic M, Valstar M, Rademaker R, Maat L (2005) Web-based database for facial expression analysis. In: Proceedings of the 13th ACM international conference on multimedia
36. Platt J (1998) Fast training of support vector machines using sequential minimal optimization. In: Advances in Kernel methods - support vector learning. MIT Press
37. Pu X, Fan K, Chen X, Ji L, Zhou Z (2015) Facial expression recognition from image sequences using twofold random forest classifier. Neurocomputing 168:1173–1180

38. Quinlan JR (1993) C4.5: programs for machine learning. Morgan Kaufmann, San Mateo
39. Sadeghi H, Raie A, Mohammadi M (2014) Facial expression recognition using texture description of displacement image. J Inf Syst Telecommun 2(4):205–212
40. Samad R, Sawada H (2011) Extraction of the minimum number of Gabor wavelet parameters for the recognition of natural facial expressions. Artif Life Robot 16(1):21–31
41. Sánchez A, Ruiz JV, Moreno AB, Montemayor AS, Hernández J, Pantrigo JJ (2011) Differential optical flow applied to automatic facial expression recognition. Neurocomputing 74(8):1272–1282
42. Shan C, Gong S, McOwan PW (2005) Robust facial expression recognition using local binary patterns. In: IEEE International conference on image processing 2005, vol 2
43. Shan C, Gong S, McOwan PW (2009) Facial expression recognition based on local binary patterns: a comprehensive study. Image Vis Comput 27(6):803–816
44. Shi J, Tomasi C (1994) Good features to track. In: 1994 Proceedings of IEEE conference on computer vision and pattern recognition, pp 593–600
45. Shin G, Chun J (2008) Spatio-temporal facial expression recognition using optical flow and HMM. In: Studies in computational intelligence, vol 149. Springer, Berlin, pp 27–38
46. Su M, Hsieh Y, Huang D (2007) A simple approach to facial expression recognition. In: CEA'07 Proceedings of the international conference on computer engineering and applications. World Scientific and Engineering Academy and Society (WSEAS), pp 456-461
47. Suk M, Prabhakaran B (2014) Real-time mobile facial expression recognition system – a case study. In: 2014 IEEE Conference on computer vision and pattern recognition workshops, pp 132–137
48. Tang J, Alelyani S, Liu H (2014) Feature selection for classification: a review, Data Classification chapter 2. CRC, Chapman & Hall
49. Tian Y, Kanade T, Cohn JF (2001) Recognizing action units for facial expression analysis. IEEE Trans Pattern Anal Mach Intell 23(2):97–115
50. Ting KC, Tan J (2013) Face recognition by neural network using Bit-Planes extracted from an image. Inf Comput Sci 10(16):5253–5261
51. Ting KC, Bong DBL, Wang YC (2008) Performance analysis of single and combined bit-planes feature extraction for recognition in face expression database. In: 2008 International conference on computer and communication engineering, pp 792–795
52. Vapnik V (1995) The nature of statistical learning theory. Springer-Verlag New York, Inc, New York
53. Viola P, Jones M (2001) Rapid object detection using a boosted cascade of simple features. In: Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition (CVPR), vol 1, pp 511–518
54. Wan S, Aggarwal JK (2014) Spontaneous facial expression recognition: a robust metric learning approach. Pattern Recogn 47(5):1859–1868
55. Wang J, Wang S, Ji Q (2014) Early facial expression recognition using Hidden Markov Models. In: 2014 22nd International conference on pattern recognition, pp 4594–4599
56. Wang Y, Yu H, Stevens B, Liu H (2015) Dynamic facial expression recognition using local patch and LBP-TOP. In: 2015 8th International conference on human system interaction (HSI), pp 362–367
57. Whitney AW (1971) A direct method of nonparametric measurement selection. IEEE Trans Comput C-20(9):1100–1103
58. Yeasin M, Bullot B (2005) Comparison of linear and non-linear data projection techniques in recognizing universal facial expressions. In: Proceedings. 2005 IEEE international joint conference on neural networks, 2005, vol 5, pp 3087–3092
59. Yu K, Wang Z, Guan G, Wu Q, Chi Z, Feng D (2012) How many frames does facial expression recognition require? In: 2012 IEEE International conference on multimedia and expo workshops, pp 290–295
60. Zhang L, Tjondronegoro D, Chandran V (2014) Facial expression recognition experiments with data from television broadcasts and the World Wide Web. Image Vis Comput 32(2):107–119
61. Zhang X, Mahoor MH, Mavadati SM (2015) Facial expression recognition using l_p-norm MKL multiclass-SVM. Mach Vis Appl 26(4):467–483
62. Zhao G, Pietikäinen M (2007) Dynamic Texture Recognition using Local Binary Patterns with an application to facial expressions. IEEE Trans Pattern Anal Mach Intell 29(6):915–928
63. Zhao G, Pietikäinen M (2009) Boosted multi-resolution spatiotemporal descriptors for facial expression recognition. Pattern Recogn Lett 30(12):1117–1127. Image/video-based Pattern Analysis and HCI Applications

Taoufik Ben Abdallah obtained his MA degree in information systems and new technologies from the Faculty of Economics and Management of Sfax-Tunisia in 2013. He is currently a PhD student at the University of Sfax, Tunisia and a member of Multimedia, InfoRmation systems and Advanced Computing Laboratory (MIRACL). His research interests include machine learning, image processing and pattern recognition.

Radhouane Guermazi received his Engineering degree in Computer Science from the National Engineering School of Sfax, Tunisia in 2003, his Master Science degree in New Technologies of Dedicated Computer Systems from the National Engineering School of Sfax, Tunisia in 2005 and the PhD degree in Computer Science from the Faculty of Economics and Management of Sfax Tunisia in 2011. In 2008, he joined the Sfax University (USS), where he was an Assistant Professor of Computer Science in Higher Institute of Computer Science and Multimedia of Sfax. He was also the chair of the master science program in Multimedia Engineering from 2012 to 2015. In 2015, he joined Saudi Electronic University where he is currently Assistant Professor of Computer Science. He is also a research member in Multimedia, InfoRmation systems and Advanced Computing Laboratory (MIRACL). His research interests include but are not limited to imbalanced data mining and pattern recognition.

Mohamed Hammami received a PhD in computer science from Ecole Centrale at the Lyon Research Center for Images and Intelligent Information Systems associated to the French research institution CNRS as UMR5205. He is currently associate professor in the Computer Science Department at the Faculty of Science Sfax-Tunisia. He is a researcher in the MIRACL-Laboratory. His current research interests in the area of computer vision and pattern recognition with a focus on biometrics and video surveillance applications. He has served as reviewer in many journals and many international conferences and workshops.