Framework for reliable, real-time facial expression recognition for low resolution images

Rizwan Ahmed Khan, Alexandre Meyer, Hubert Konik, Saïda Bouakaz

Framework for reliable, real-time facial expression recognition for low resolution images

Rizwan Ahmed Khan[a,b,*], Alexandre Meyer[a,b], Hubert Konik[a,c], Saida Bouakaz[a,b]

[a] *Université de Lyon, CNRS*
[b] *Université Lyon 1, LIRIS, UMR5205, F-69622, France*
[c] *Université Jean Monnet, Laboratoire Hubert Curien, UMR5516, 42000 Saint-Etienne, France*

Abstract

Automatic recognition of facial expressions is a challenging problem specially for low spatial resolution facial images. It has many potential applications in human-computer interactions, social robots, deceit detection, interactive video and behavior monitoring. In this study we present a novel framework that can recognize facial expressions very efficiently and with high accuracy even for very low resolution facial images. The proposed framework is memory and time efficient as it extracts texture features in a pyramidal fashion only from the perceptual salient regions of the face. We tested the framework on different databases, which includes Cohn-Kanade (CK+) posed facial expression database, spontaneous expressions of MMI facial expression database and FG-NET facial expressions and emotions database (FEED)

*corresponding author for this article is Rizwan A Khan, Phone: (+33) (0)4 72 43 19 75, Cell: (+33) (0) 63 86 54 278, Fax: (+33) (0)4 72 43 15 36

Email addresses: `Rizwan-ahmed.khan@liris.cnrs.fr` (Rizwan Ahmed Khan), `Alexandre.meyer@liris.cnrs.fr` (Alexandre Meyer), `Hubert.Konik@univ-st-etienne.fr` (Hubert Konik), `Saida.bouakaz@liris.cnrs.fr` (Saida Bouakaz)

and obtained very good results. Moreover, our proposed framework exceeds state-of-the-art methods for expression recognition on low resolution images.

Keywords:
Facial expression recognition, Low Resolution Images, Local Binary Pattern, Image pyramid, Salient facial regions

---

1. Introduction

Communication in any form i.e. verbal or non-verbal is vital to complete various routine tasks and plays a significant role in daily life. Facial expression is the most effective form of non-verbal communication and it provides a clue about emotional state, mindset and intention [7]. Human visual system (HVS) decodes and analyzes facial expressions in real time despite having limited neural resources. As an explanation for such performance, it has been proposed that only some visual inputs are selected by considering "salient regions" [36], where "salient" means most noticeable or most important.

For computer vision community it is a difficult task to automatically recognize facial expressions in real-time with hight reliability. Variability in pose, illumination and the way people show expressions across cultures are some of the parameters that make this task difficult. Low resolution input images makes this task even harder. Smart meeting, video conferencing and visual surveillance are some of the real world applications that require facial expression recognition system that works adequately on low resolution images. Another problem that hinders the development of such system for real world application is the lack of databases with natural displays of expressions [27]. There are number of publicly available benchmark databases with posed displays of the six basic emotions [6] exist but there is no equivalent of this for spontaneous basic emotions. While, It has been proved that Spontaneous facial expressions differ substantially from posed expressions [2]. In this work, we propose a facial expression recognition system that caters for illumination changes and works equally well for low resolution as well as for good quality / high resolution images. We have tested our proposed system on spontaneous facial expressions as well and recorded encouraging results.

We propose a novel descriptor for facial features analysis, Pyramid of Local Binary Pattern (PLBP) (refer Section 3). PLBP is a spatial representation of local binary pattern (LBP) [19] and it represents stimuli by its local texture (LBP) and the spatial layout of the texture. We combined pyramidal approach with LBP descriptor for facial feature analysis as this approach has already been proved to be very effective in a variety of image processing tasks [10]. Thus, the proposed descriptor is a simple and computationally efficient extension of LBP image representation, and it shows significantly improved performance for facial expression recognition tasks for low resolution images. We base our framework for automatic facial expression recognition (FER) on human visual system (HVS) (refer Section 5), so it extracts PLBP features only from the salient regions of the face. To determine which facial region(s) is the most important or salient according to HVS, we conducted a psycho-visual experiment using an eye-tracker (refer Section 4). We considered six universal facial expressions for psycho-visual experimental study as these expressions are proved to be consistent across cultures [6]. These six expressions are anger, disgust, fear, happiness, sadness and surprise. The novelty of the proposed framework is that, it is illumination invariant, reliable on low resolution images and works adequately for both i.e. posed and spontaneous expressions.

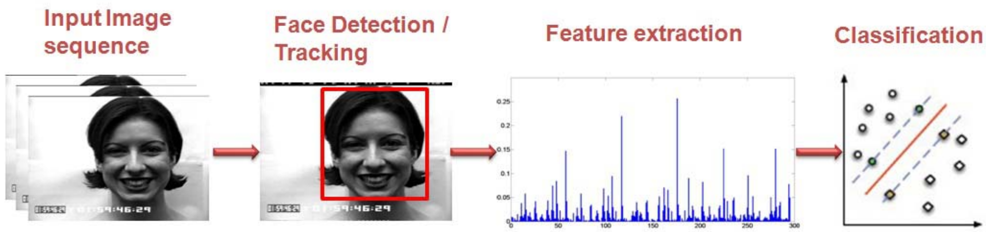

Figure 1: Basic structure of facial expression recognition system pipeline.

Generally, facial expression recognition system consists of three steps: face detection, feature extraction and expression classification. The same has been shown in Figure 1. In our framework we tracked face / salient facial regions using Viola-Jones object detection algorithm [30] as it is the most cited and considered the fastest and most accurate pattern recognition method for face detection [13]. The second step in the framework is feature extraction, which is the area where this study contributes. The optimal features should minimize within-class variations of expressions, while maximize between class variations. If inadequate features are used, even the best classifier could fail to achieve accurate recognition [25]. Section 3 presents the novel method for facial features extraction which is based on human visual system (HVS). To study and understand HVS we performed psycho-visual experiment. Psycho-visual experimental study is briefly described in Section 4. Expression classification or recognition is the last step in the pipeline. In literature two different ways are prevalent to recognize expressions i.e. direct recognition of prototypic expressions or recognition of expressions through facial action coding system (FACS) action units (AUs) [8]. In our proposed framework, which is described in Section 5 we directly classify six universal prototypic expressions [6]. The performance of the framework is evaluated for five different classifiers (from different families i.e. classification Tree, Instance Based Learning, SVM etc) and results are presented in Section 6. Next section presents the brief literature review for facial features extraction methods.

2. Related work

In the literature, various methods are employed to extract facial features and these methods can be categorized either as appearance-based methods or geometric feature-based methods.

Appearance-based methods. One of the widely studied method to extract appearance information is based on Gabor wavelets [15, 26, 5]. Generally, the drawback of using Gabor filters is that it produces extremely large number of features and it is both time and memory intensive to convolve face images with a bank of Gabor filters to extract multi-scale and multi-orientational coefficients. Another promising approach to extract appearance information is by using Haar-like features, see Yang et al. [33]. Recently, texture descriptors and classification methods i.e. Local Binary Pattern (LBP) [19] and Local Phase Quantization (LPQ) [21] are also studied to extract appearance-based facial features. Zhao et al. [35] proposed to model texture using volume local binary patterns (VLBP) an extension to LBP, for expression recognition.

Geometric-based methods. Geometric feature-based methods [34, 22, 28, 1] extracts shapes and locations of facial components information to form a feature vector. The problem with using geometric feature-based methods is that they usually require accurate and reliable facial feature detection and tracking which is difficult to achieve in many real world applications where illumination changes with time and images are recorded in very low resolution.

Generally, we have found that all the reviewed methods for automatic facial expression recognition are computationally expensive and usually requires dimensionally large feature vector to complete the task. This explains their inability for real-time applications. Secondly, in literature, very few studies exist that tackles the issue of expressions recognition from low resolution images, this adds to lack of applicability of expression recognition system for real world applications. Lastly, all of the reviewed methods, spend computational time on whole face image or divides the facial image based on some mathematical or geometrical heuristic for features extraction. We argue that the task of expression analysis and recognition could be done in more conducive manner, if only some regions are selected for further processing (i.e. salient regions) as it happens in human visual system. Thus, our contributions in this study are:

1. We propose a novel descriptor for facial expression analysis i.e. Pyramid of Local Binary Pattern (PLBP), which outperforms state-of-the-art methods for expression recognition on low resolution images (spatially degraded images). It also performs better than other state-of-the-art methods for good resolution images (with no degradation).

2. As the proposed framework is based on human visual system it algorithmically processes only salient facial regions which reduces the length of feature vector. This reduction in feature vector length makes the proposed framework suitable for real-time applications due to minimized computational complexity.

3. Pyramid of Local Binary Pattern

The proposed framework creates a novel feature space by extracting proposed PLBP (pyramid of local binary pattern) features only from the visually salient facial region (see Section 4 for psycho-visual experiment). PLBP is a *pyramidal-based spatial* representation of local binary pattern (LBP) descriptor. PLBP represents stimuli by their local texture (LBP) and the spatial layout of the texture. The spatial layout is acquired by tiling the image into regions at multiple resolutions. The idea is illustrated in Figure 2. If only the coarsest level is used, then the descriptor reduces to a global LBP histogram. Comparing to the multi-resolution LBP of Ojala et al.[20] , our descriptor selects samples in a more uniformly distributed manner, whereas Ojala's LBP takes samples centered around a point leading to missing some information in the case of face (which is different than a repetitive texture).

LBP features were initially proposed for texture analysis [19], but recently they have been successfully used for facial expression analysis [35, 25]. The most important property of LBP features are their tolerance against illumination changes and their computational simplicity [18, 19, 20]. The operator labels the pixels of an image by thresholding the 3 x 3 neighbourhood of each pixel with the center value and considering the result as a binary num-

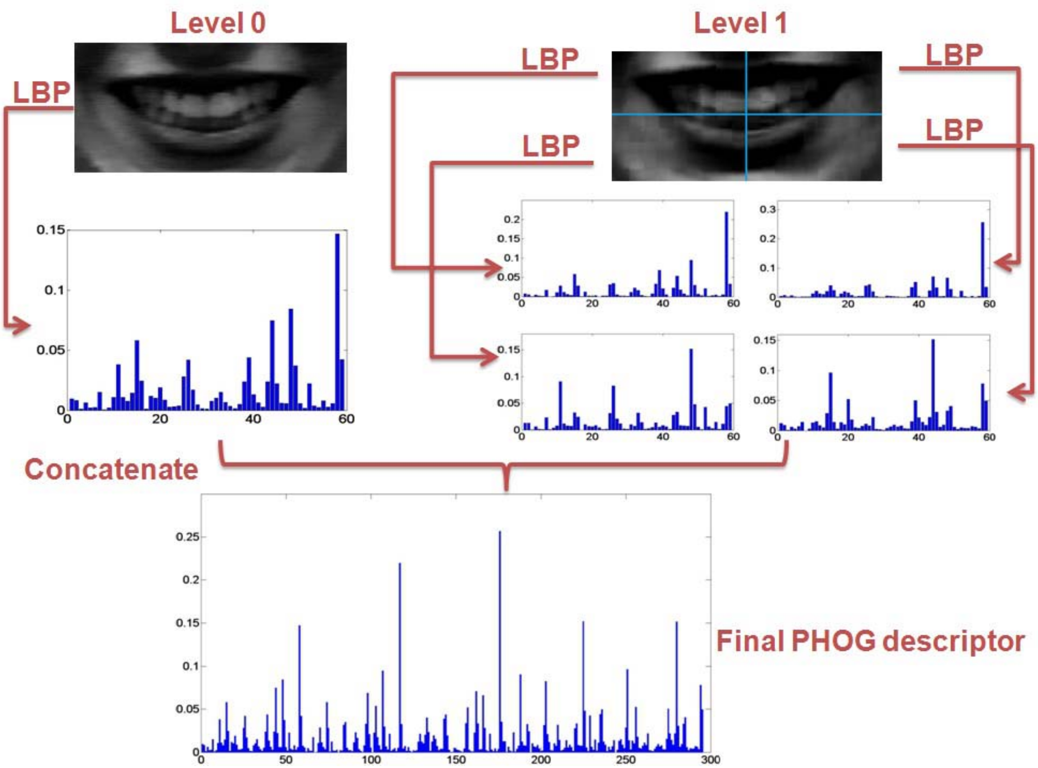

Figure 2: Pyramid of Local Binary Pattern. First row: stimuli at two different pyramid levels, second row: histograms of LBP at two respective levels, third row: final descriptor.

ber. Then the histogram of the labels can be used as a texture descriptor. Formally, LBP operator takes the form:

$$LBP(x_c, y_c) = \sum_{n=0}^{7} s(i_n - i_c)2^n \tag{1}$$

where in this case n runs over the 8 neighbours of the central pixel c, i_c and i_n are the grey level values at c and n and $s(u)$ is 1 if $u \geq 0$ or 0 otherwise.

Later, the LBP operator is extended to use neighborhood of different sizes

[20] as the original operator uses 3 x 3 neighbourhood. Using circular neighborhoods and bilinearly interpolating the pixel values allow any radius and number of pixels in the neighborhood. The LBP operator with P sampling points on a circular neighborhood of radius R is given by:

$$LBP_{P,R} = \sum_{p=0}^{P-1} s(g_p - g_c)2^p \tag{2}$$

Another extension to the original operator is the definition of *uniform patterns*, which can be used to reduce the length of the feature vector and implement a simple rotation-invariant descriptor. A local binary pattern is called uniform if the binary pattern contains at most two bitwise transitions from 0 to 1 or vice versa when the bit pattern is traversed circularly. Accumulating the patterns which have more than 2 transitions into a single bin yields an LBP operator, denoted $LBP^{u2}_{P,R}$. patterns. These binary patterns can be used to represent texture primitives such as spot, flat area, edge and corner.

We extend LBP operator so that the stimuli can be represented by its local texture and the spatial layout of the texture. We call this extended LBP operator as pyramid of local binary pattern or PLBP. PLBP creates the spatial pyramid by dividing the stimuli into finer spatial sub-regions by iteratively doubling the number of divisions in each dimension. It can be observed from the Figure 2 that the pyramid at level l has 2^l sub-regions along each dimension $(R_0, \dots R_m)$. Histograms of LBP features at the same levels are concatenated. Then, their concatenation at different pyramid levels gives final PHOG descriptor (as shown in Figure 2). It can be defined as:

$$H_{i,j} = \sum_{l} \sum_{xy} I\{f_l(x,y) = i\} I\{(x,y) \in R_l\} \tag{3}$$

where $l = 0 \ldots m-1$, $i = 0 \ldots n-1$. n is the number of different labels produced by the LBP operator and

$$I(A) = \begin{cases} 1 & \text{if A is true ,} \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

While, the dimensionality of the descriptor can be calculated by:

$$N \sum_{l} 4^l \tag{5}$$

Where, in our experiment (see Section 6) l=1 and N= 59 as we created pyramid up to level 1 and extracted 59 LBP features using $LBP^{u2}_{8,2}$ operator, which denotes a uniform LBP operator with 8 sampling pixels in a local neighborhood region of radius 2. This pattern reduces the histogram from 256 to 59 bins. In our experiment we obtained 295 dimensional feature vector from one facial region i.e. mouth region (59 dimensions / sub-region), since we executed the experiment with the pyramid of level 1 (the same is shown in Figure 2).

3.1. Novelty of the proposed descriptor

There exist some methods in literature that uses Pyramid of LBP for different applications and they look similar to our proposed descriptor i.e. [32, 9, 17]. Our proposition is novel and there exist differences in the methodology that creates differences in the extracted information. Method for face recognition proposed in [32] creates pyramids before applying LBP operator by down sampling original image i.e. scale-space representation, whereas we propose to create the spatial pyramid by dividing the stimuli into finer spatial sub-regions by iteratively doubling the number of divisions in each dimension. Secondly, our approach reduces memory consumption (do not requires to store same image in different resolutions) and is computationally more efficient. Guo et al. [9] proposed approach for face and palmprint recognition based on multiscale LBP. Their proposed method seems similar to our method for expression recognition but how multiscale analysis is achieved deviates our approach. Approach proposed in [9] achieves multiscale analysis using different values of P and R, where $LBP(P, R)$ denotes a neighborhood of P equally spaced sampling points on a circle of radius R (discussed earlier). Same approach has been applied by Moore et al. [17] for facial features analysis. Generally the drawback of using such approach is that it increases the size of the feature histogram and increases the computational cost. [17] reports dimensionality of feature vector as high as 30,208 for multiscale face expression analysis as compared to our proposition which creates 590 dimensional feature vector (see Section 5) for the same task. We achieve the task of multiscale analysis much more efficiently than any other earlier proposed methods. By the virtue of efficient multiscale analysis our framework can be used for real time applications (see Table 1 for the time and memory consumption comparison) which is not the case with other methods.

As mentioned earlier, we base our framework for facial expression recognition on human visual system (HVS), which selects only few facial regions (salient) to extract information. In order to determine the saliency of facial region(s) for a particular expression, we conducted psycho-visual experiment with the help of an eye-tracker. Next section briefly explains the psycho-visual experimental study.

4. Psycho-Visual experiment

The aim of our experiment was to record the eye movement data of human observers in free viewing conditions. The data were analyzed in order to find which components of face are salient for specific displayed expression.

4.1. Participants, apparatus and stimuli

Eye movements of fifteen human observers were recorded using video based eye-tracker (EyelinkII system, SR Research), as the subjects watched the collection of 54 videos selected from the extended Cohn-Kanade (CK+) database [16], showing one of the six universal facial expressions [6]. Observers include both male and female aging from 20 to 45 years with normal or corrected to normal vision. All the observers were naïve to the purpose of an experiment.

4.2. Eye movement recording

Eye position was tracked at 500 Hz with an average noise less than 0.01°. Head mounted eye-tracker allows flexibility to perform the experiment in free viewing conditions as the system is designed to compensate for small head movements.

4.3. Psycho-Visual experiment Results

In order to statistically quantify which region is perceptually more attractive for specific expression, we have calculated the average percentage of trial

Figure 3: Summary of the facial regions that emerged as salient for six universal expressions. Salient regions are mentioned according to their importance (for example facial expression of "fear" has two salient regions but mouth is the most important region according to HVS).

time observers have fixated their gazes at specific region(s) in a particular time period. As the stimuli used for the experiment is dynamic i.e. video sequences, it would have been incorrect to average all the fixations recorded during trial time (run length of the video) for the data analysis as this could lead to biased analysis of the data. To meaningfully observe and analyze the gaze trend across one video sequence we have divided each video sequence in three mutually exclusive time periods. The first time period correspond to initial frames of the video sequence i.e. neutral face. The last time period encapsulates the frames where the expression is shown with full intensity (apex frames). The second time period is a encapsulation of the frames which has a transition of facial expression i.e. transition from neutral face to the beginning of the desired expression (i.e neutral to the onset of the expression). Then the fixations recorded for a particular time period are averaged across fifteen observers. For drawing the conclusions we considered second and third time periods as they have the most significant information in terms of specific displayed expression. Conclusions drawn are summarized in Figure 3. Refer [11] for the detailed explanation of the psycho-visual experimental study.

5. Expression Recognition Framework

Feature selection along with the region(s) from where these features are going to be extracted is one of the most important step to recognize expressions. As the proposed framework draws its inspiration from the human visual system (HVS), it extracts proposed features i.e. PLBP, only from the perceptual salient facial region(s) which were determined through Psycho-Visual experiment. Schematic overview of the framework is presented in Figure. 4. Steps of the proposed framework are as follows:

1. First, the framework extracts PLBP features from the mouth region, feature vector of 295 dimensions ($f_1, ..., f_{295}$). The classification ("Classifier-a" in the Figure. 4) is carried out on the basis of extracted features in order to make two groups of facial expressions. First group comprises of those expressions that has one perceptual salient region i.e. happiness, sadness and surprise while the second group is composed of those expressions that have two or more perceptual salient regions i.e.

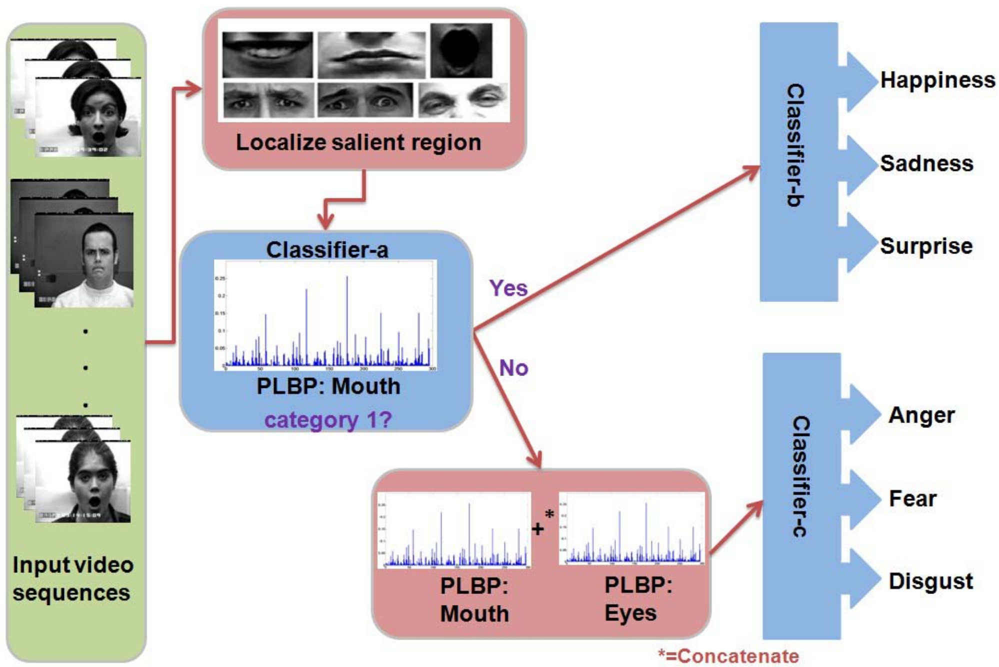

Figure 4: Schematic overview of the framework.

anger, fear and disgust (see Section 4.3). Purpose of making two groups of expressions is to reduce feature extraction computational time.

2. If the stimuli is classified in the first group, then it is classified either as happiness, sadness or surprise by the "Classifier-b" using already extracted PLBP features from the mouth region.

3. If the stimuli is classified in the second group, then the framework extracts PLBP features from the eyes region (it is worth mentioning here that for the expression of "disgust" nose region emerged as the salient but the framework do not explicitly extracts features from the nose region as the region of nose that emrged as salient is the upper nose wrinkle area which is connected and already included in the localiza-

tion of the eyes region, refer Figure 3) and concatenates them with the already extracted PLBP features from the mouth region, feature vector of 590 dimensions ($f_1, ..., f_{295} + f_1, ..., f_{295}$) . Then, the concatenated feature vector is fed to the classifier ("Classifier-c") for the final classification.

6. Experiment and results

We performed person-independent facial expression recognition using proposed PLBP features [1]. We performed four experiments to test different scenarios.

1. First experiment was performed on the extended Cohn-Kanade (CK+) database [16]. This database contains 593 sequences of posed universal expressions.
2. Second experiment was performed to test the performance of the proposed framework on low resolution image sequences.
3. Third experiment tests the robustness of the proposed framework when generalizing on the new dataset.
4. Fourth experiment was performed on the MMI facial expression database (Part IV and V of the database) [27] which contains spontaneous/natural expressions.

For the first two experiments we used all the 309 sequences from the CK+ database which have FACS coded expression label [8]. The experiment

[1] video showing the result of the proposed framework on good quality image sequences is available at: http://liris.cnrs.fr/~rakhan/FER_Demo.wmv

was carried out on the frames which covers the status of onset to apex of the expression, as done by Yang et al. [33]. Region of interest was obtained automatically by using Viola-Jones object detection algorithm [30] and processed to obtain PLBP feature vector. We extracted LBP features only from the salient region(s) using $LBP_{8,2}^{u2}$ operator which denotes a uniform LBP operator with 8 sampling pixels in a local neighborhood region of radius 2. Only exception was in the second experiment, when we adopted $LBP_{4,1}^{u2}$ operator when the spatial facial resolution gets smaller than 36 x 48.

In our framework we created image pyramid up to level 1, so in turn got five sub-regions from one facial region i.e. mouth region (see Figure. 2). In total we obtained 295 dimensional feature vector (59 dimensions / sub-region). As mentioned earlier we adopted $LBP_{4,1}^{u2}$ operator when the spatial facial resolution was 18 x 24. In this case we obtained 75 dimensional feature vector (15 dimensions / sub-region).

We recorded correct classification accuracy in the range of 95% for image pyramid level 1. We decided not to test framework with further image pyramid levels as it would double the size of feature vector and thus increase the feature extraction time and likely would add few percents in the accuracy of the framework which will be insignificant for a framework holistically.

6.1. First experiment: posed expressions

This experiment measures the performance of the proposed framework on the classical database i.e. extended Cohn-Kanade (CK+) database [16]. Most of the methods in literature report their performance on this database, so this experiment could be considered as the benchmark experiment for facial expression recognition framework.

The performance of the framework was evaluated for five different classifiers:

1. Support Vector Machine (SVM)" with χ^2 kernel and γ=1
2. C4.5 Decision Tree (DT) with reduced-error pruning
3. Random Forest (RF) of 10 trees
4. *2* Nearest Neighbor (*2*NN) based on Euclidean distance
5. Naive Bayes (NB) classifier

Above mentioned classifiers are briefly described below.

Support vector machine (SVM). SVM performs an implicit mapping of data into a higher dimensional feature space, and then finds a linear separating hyperplane with the maximal margin to separate data in this higher dimensional space [29]. Given a training set of labeled examples $\{ (x_i, y_i) , i = 1 \ldots l \}$ where $x_i \in \Re^n$ and $y_i \in$ {-1, 1}, a new test example x is classified by the following function:

$$f(x) = sgn(\sum_{i=1}^{l} \alpha_i y_i K(x_i, x) + b) \tag{6}$$

where α_i are Langrange multipliers of a dual optimization problem that describe the separating hyperplane, $K(.,.)$ is a kernel function, and b is the threshold parameter of the hyperplane. We used Chi-Square kernel as it is best suited for histograms. It is given by:

$$K(x, y) = 1 - \sum_i \frac{2 \times (x_i - y_i)^2}{(x_i + y_i)} \tag{7}$$

Classification Trees. A Classification Tree is a classifier composed by nodes and branches which break the set of samples into a set of covering decision rules. In each node, a single test is made to obtain the partition. The starting node is called the root of the tree. In the final nodes or leaves, a decision about the classification of the case is made. In this work, we have used C4.5 paradigm [24]. Random Forest (RFs) are collections of Decision Trees (DTs) that have been constructed randomly. RFs generally performs better than DT on unseen data.

Instance Based Learning. k-NN classifiers are instance-based algorithms taking a conceptually straightforward approach to approximating real or discrete valued target functions. The learning process consists in simply storing the presented data. All instances correspond to points in an n-dimensional space and the nearest neighbors of a given query are defined in terms of the standard Euclidean distance. The probability of a query q belonging to a class c can be calculated as follows:

$$p(c \mid q) = \frac{\sum\limits_{k \in K} W_k.1_{(kc=c)}}{\sum\limits_{k \in K} W_k} \tag{8}$$

$$W_k = \frac{1}{d(k,q)} \tag{9}$$

K is the set of nearest neighbors, kc the class of k and $d(k,q)$ the Euclidean distance of k from q.

Naive Bayes Classifiers. The Naive-Bayes (NB) classifier uses the Bayes theorem to predict the class for each case, assuming that the predictive genes are independent given the category. To classify a new sample characterized by d genes $X = (X1, X2, \ldots Xd)$, the NB classifier applies the following rule:

$$C_N - B = arg\, max_{c_j \in C}\, p(c_j) \prod_{i=1}^{d} p(x_i \mid c_j) \tag{10}$$

where $C_N - B$ denotes the class label predicted by the Naive-Bayes classifier and the possible classes of the problem are grouped in $C = \{c_1, \ldots c_l\}$.

6.1.1. Results

The framework achieved average recognition rate of 96.7%, 97.9%, 96.2%, 94.7 % and 90.2 % for SVM, *2*Nearest Neighbor (*2*NN), Random Forest (RF), C4.5 Decision Tree (DT) and Naive Bayes (NB) respectively using *10*-fold cross validation technique. One of the most interesting aspects of our approach is that it gives excellent results for a simple *2*NN classifier which is a non-parametric method. This points to the fact that framework do not need computationally expensive methods such as SVM, random forests or decision trees to obtain good results. In general, the proposed framework achieved high expression recognition accuracies irrespective of the classifiers, proves the descriptive strength of the extracted features (features minimizes within-class variations of expressions, while maximizes between class variations). For comparison and reporting results, we have used the classification results obtained by the SVM as it is the most cited method for classification in the literature.

6.1.2. Comparisons

We chose to compare average recognition performance of our framework with the framework proposed by Shan et.al [25] with different SVM kernels.

Table 1: Comparison of time and memory consumption.

	LBP [25]	Gabor [25]	Gabor [3]	**PLBP**
Memory (feature dimension)	2,478	42,650	92,160	**590**
Time(feature extraction time)	0.03s	30s	-	**0.01s**

Our choice was based on the fact that both have common underlying descriptor i.e. local binary pattern (LBP), secondly framework proposed by Shan et.al [25] is highly cited in the literature. Our framework obtained average recognition percentage of 93.5% for SVM linear kernel while for the same kernel Shan et.al [25] have reported 91.5%. For SVM with polynomial kernel and SVM with RBF kernel our framework achieved recognition accuracy of 94.7% and 94.9% respectively, as compared to 91.5% and 92.6%.

In terms of time and memory costs of feature extraction process, we have measured and compared our descriptor with the LBP and Gabor-wavelet features in Table 1. Table 1 shows the effectiveness of the proposed descriptor for facial feature analysis i.e. PLBP, for real-time applications as it is memory efficient and its extraction time is much lower than other compared descriptor (see Section 5 for the dimensionality calculation). In Table 1 feature dimension reported are stored in a data type "float" and float occupies four bytes. The proposed framework is compared with other state-of the-art frameworks using same database (i.e Cohn-Kanade database) and the results are presented in Table 2.

Table 2 shows the comparison of the achieved average recognition rate of the proposed framework with the state-of-the-art methods using same database (i.e Cohn-Kanade database). Results from [33] are presented for the

Table 2: Comparison with the state-of-the-art methods for posed expressions.

	Sequence Num	Class Num	Performance Measure	Recog. Rate (%)
[15]	313	7	leave-one-out	93.3
[35]	374	6	2-fold	95.19
[35]	374	6	10-fold	96.26
[14]	374	6	5-fold	94.5
[26]	375	6	-	93.8
[33]a	352	6	66% split	92.3
[33]b	352	6	66% split	80
Ours	**309**	**6**	**10-fold**	**96.7**
Ours	**309**	**6**	**2-fold**	**95.2**

two configurations. "[33]a" shows the result when the method was evaluated for the last three frames from the sequence while "[33]b" presents the reported result for the frames which encompasses the status from onset to apex of the expression. It can be observed from the Table 2 that the proposed framework is comparable to any other state-of-the-art method in terms of expression recognition accuracy. The method discussed in "[33]b" is directly comparable to our method, as we also evaluated the framework on similar frames. In this configuration, our framework is better in terms of average recognition accuracy.

In general, Table 1 and 2 show that the framework is better than the state-of-the-art frameworks in terms of average expression recognition performance, time and memory costs of feature extraction processes. These results show that the system could be used with the high degree of confidence for real-time applications as its unoptimized Matlab implementation runs at more than 30 frames/second (30 fps).

6.2. Second experiment: low resolution image sequences

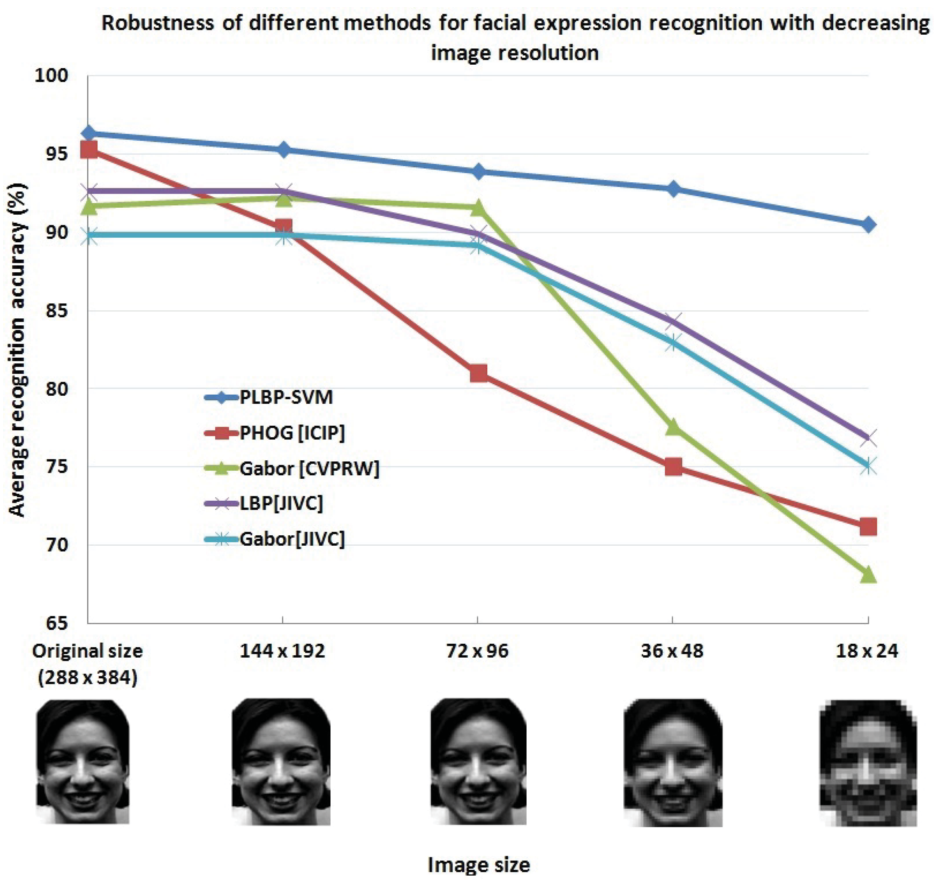

Figure 5: Robustness of different methods for facial expression recognition with decreasing image resolution. PHOG[ICIP] corresponds to framework proposed by Khan et. al [12], Gabor [CVPRW] corresponds to Tian's work [26], LBP[JIVC] and Gabor[JIVC] corresponds to results reported by Shan et. al [25]

Most of the existing state-of-the-art systems for expressions recognition report their results on high resolution images with out reporting results on low resolution images. As mentioned earlier there are many real world applications that require expression recognition system to work amicably on low resolution images. Smart meeting, video conferencing and visual surveillance are some examples of such applications. To compare with Tian's work [26], we tested our proposed framework on low resolution images of four different facial resolutions (144 x 192, 72 x 96, 36 x 48, 18 x 24) based on Cohn-Kanade database. Tian's work can be considered as the pioneering work for low resolution image facial expression recognition. Figure 5 shows the images at different spatial resolution along with the average recognition accuracy achieved by the different methods. Low resolution image sequences were obtained by down sampling the original sequences. All the other experimental parameters i.e. descriptor, number of sequences and region of interest, were same as mentioned earlier in the Section 6.

Figure 5 reports the recognition results of the proposed framework with the state-of-the-art methods on four different low facial resolution images. Reported results of our proposed method i.e. are obtained using support vector machine (SVM)" with χ^2 kernel and γ=1. In Figure 5 recognition curve for our proposed method is shown as *PLBP-SVM*, recognition curves of LBP [25] and Gabor [25] are shown as *LBP[JIVC]* and *Gabor[JIVC]* respectively, curve for Tian's work [26] is shown as *Gabor[CVPRW]* while Khan et al. [12] proposed system's curve is shown as *PHOG[ICIP]*. Results reports in LBP [25] and Gabor [25], the different facial image resolution are 110 x 150, 55 x 75, 27 x 37 and 14 x 19 which are comparable to the resolutions of 144 x 192, 72 x 96, 36 x 48, 18 x 24 pixels in our experiment. Referenced figure shows the supremacy of the proposed framework for low resolution images. Specially for the smallest tested facial image resolution (18 x 24) our framework performs much better than any other compared state-of-the-art method.

Results from the first and second experiment show that the proposed framework for facial expression recognition works amicably on classical dataset (CK dataset) and its performance is not effected significantly for low resolution images. Secondly, the framework has a very low memory requirement and thus it can be utilized for real-time applications.

6.3. Third experiment: generalization on the new dataset

The aim of this experiment is to study how well the proposed framework generalizes on the new dataset. We used image sequences from CK+ dataset and FG-NET FEED (Facial Expressions and Emotion Database) [31]. FG-NET FEED contains 399 video sequences across 18 different individuals showing seven facial expressions i.e. six universal expression [6] plus one neutral. In this dataset individuals were not asked to act rather expressions were captured while showing them video clips or still images.

The experiment was carried out on the frames which covers the status of onset to apex of the expression as done in the previous experiment. This experiment was performed in two different scenarios, with the same classifier parameters as the first experiment:

a. In the first scenario samples from the CK+ database were used for the training of different classifiers and samples from FG-NET FEED [31] were used for the testing. Obtained results are presented in Table 3.

b. In the second scenario we used samples from the FG-NET FEED for the training and testing was carried out with the CK+ database samples. Results obtained are presented in last two rows of Table 3.

This experiment simulates the real life situation when the framework would be employed to recognize facial expressions on the unseen data. Obtained results are presented in Table 3. Reported average recognition percentages for training phase were calculated using *10*-fold cross validation method. Obtained results are encouraging and they can be further improved by training classifiers on more than one dataset before using in real life scenario.

Table 3: Average recognition accuracy (%)

Training on CK+ database and testing it with FG-NET FEED				
	SVM	C4.5 DT	RF	2NN
Training samples	96.7	94.7	96.2	97.9
Test samples	81.9	74.8	79.5	83.1
Training on FG-NET FEED and testing it with CK+ database				
Training samples	92.3	91.2	90.5	93.3
Test samples	80.5	77.3	79	84.7

6.4. Fourth experiment: spontaneous expressions

Spontaneous/natural facial expressions differ substantially from posed expressions [2]. The same has also been proved by psychophysical work [7]. To test the performance of the proposed framework on the spontaneous facial expressions we used 392 video segments from part IV and V of the MMI facial expression database [27]. Part IV and V of the database contains spontaneous/naturalistic expressions recorded from 25 participants aged between 20 and 32 years in two different settings. Due to ethical concerns the database contains only the video recording of the expressions of happiness, surprise and disgust [27].

The framework achieved average recognition rate of 91%, 91.4%, 90.3% and 88% for SVM, 2-nearest neighbor, Random forest and C4.5 decision tree respectively using *10*-fold cross validation technique. Algorithm of Park et al.[23] for spontaneous expression recognition achieved results for three expressions in the range of 56% to 88% for four different configurations which is less than recognition rate of our proposed algorithm, although results cannot be compared directly as they used different database.

7. Conclusions and future work

We presented a novel descriptor and framework for automatic and reliable facial expression recognition. Framework is based on initial study of human vision and works adequately on posed as well as on spontaneous expressions. The key conclusion drawn from the study are:

1. Facial expressions can be analyzed automatically by mimicking human visual system i.e. extracting features only from the salient facial regions.
2. Features extracted using proposed pyramidal local binary pattern (PLBP) operator have strong discriminative ability as the recognition result for

six universal expressions is not effected by the choice of classifier.

3. The proposed framework is robust for low resolution images, spontaneous expressions and generalizes well on unseen data.
4. The proposed framework can be used for real-time applications since its unoptimized Matlab implementation run at more than 30 frames / second (30 fps) on a Windows 64 bit machine with i7 processor running at 2.4 GHz having 6 GB of RAM.

In future we plan to investigate the effect of occlusion as this parameter could significantly impact the performance of the framework for real world applications. Secondly, the notion of movement could improve the performance of the proposed framework for real world applications as the experimental study conducted by Bassili [4] suggested that dynamic information is important for facial expression recognition. Another parameter that needs to be investigated is the variations of camera angle as for many applications frontal facial pose is difficult to record.

References

[1] Bai, Y., Guo, L., Jin, L., Huang, Q., 2009. A novel feature extraction method using pyramid histogram of orientation gradients for smile recognition, in: International Conference on Image Processing.

[2] Bartlett, M.S., Littlewort, G., Braathen, B., Sejnowski, T.J., Movellan, J.R., 2002. A prototype for automatic recognition of spontaneous facial actions, in: Advances in Neural Information Processing Systems.

[3] Bartlett, M.S., Littlewort, G., Fasel, I., Movellan, J.R., 2003. Real time face detection and facial expression recognition: Development and applications to human computer interaction., in: Conference on Computer Vision and Pattern Recognition Workshop, 2003.

[4] Bassili, J.N., 1979. Emotion recognition: The role of facial movement and the relative importance of upper and lower areas of the face. Journal of Personality and Social Psychology 37, 2049–2058.

[5] Donato, G., Bartlett, M.S., Hager, J.C., Ekman, P., Sejnowski, T.J., 1999. Classifying facial actions. IEEE Transaction on Pattern Analysis and Machine Intelligence 21, 974–989.

[6] Ekman, P., 1971. Universals and cultural differences in facial expressions of emotion, in: Nebraska Symposium on Motivation, Lincoln University of Nebraska Press. pp. 207–283.

[7] Ekman, P., 2001. Telling Lies: Clues to Deceit in the Marketplace, Politics, and Marriage. W. W. Norton & Company, New York. 3rd edition.

[8] Ekman, P., Friesen, W., 1978. The facial action coding system: A technique for the measurement of facial movements. Consulting Psychologist .

[9] Guo, Z., Zhang, L., Zhang, D., Mou, X., 2010. Hierarchical multiscale lbp for face and palmprint recognition, in: IEEE International Conference on Image Processing, pp. 4521–4524.

[10] Hadjidemetriou, E., Grossberg, M., Nayar, S., 2004. Multiresolution histograms and their use for recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 26, 831 –847.

[11] Khan, R.A., Meyer, A., Konik, H., Bouakaz, S., 2012a. Exploring human visual system: study to aid the development of automatic facial expression recognition framework, in: Computer Vision and Pattern Recognition Workshop.

[12] Khan, R.A., Meyer, A., Konik, H., Bouakaz, S., 2012b. Human vision inspired framework for facial expressions recognition, in: IEEE (Ed.), IEEE International Conference on Image Processing.

[13] Kolsch, M., Turk, M., 2004. Analysis of rotational robustness of hand detection with a viola-jones detector, in: 17th International Conference on Pattern Recognition, pp. 107 – 110 Vol.3.

[14] Kotsia, I., Zafeiriou, S., Pitas, I., 2008. Texture and shape information fusion for facial expression and facial action unit recognition. Pattern Recognition 41, 833–851.

[15] Littlewort, G., Bartlett, M.S., Fasel, I., Susskind, J., Movellan, J., 2006. Dynamics of facial expression extracted automatically from video. Image and Vision Computing 24, 615–625.

[16] Lucey, P., Cohn, J.F., Kanade, T., Saragih, J., Ambadar, Z., Matthews, I., 2010. The extended cohn-kande dataset (CK+): A complete facial expression dataset for action unit and emotion-specified expression., in:

IEEE Conference on Computer Vision and Pattern Recognition Workshops.

[17] Moore, S., Bowden, R., 2011. Local binary patterns for multi-view facial expression recognition. Computer Vision and Image Understanding 115, 541 – 558.

[18] Ojala, T., Pietikäinen, M., 1999. Unsupervised texture segmentation using feature distributions. Pattern Recognition 32, 477– 486.

[19] Ojala, T., Pietikäinen, M., Harwood, D., 1996. A comparative study of texture measures with classification based on featured distribution. Pattern Recognition 29, 51–59.

[20] Ojala, T., Pietikäinen, M., Mäenpää, T., 2002. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Transaction on Pattern Analysis and Machine Intelligence 24, 971–987.

[21] Ojansivu, V., Heikkilä, J., 2008. Blur insensitive texture classification using local phase quantization, in: International conference on Image and Signal Processing.

[22] Pantic, M., Patras, I., 2006. Dynamics of facial expression: recognition of facial actions and their temporal segments from face profile image sequences. IEEE Transactions on Systems, Man, and Cybernetics 36, 433–449.

[23] Park, S., Kim, D., 2008. Spontaneous facial expression classification

with facial motion vector, in: IEEE Conference on Automatic Face and Gesture Recognition.

[24] Quinlan, J.R., 1993. C4.5: Programs for Machine Learning. Morgan Kaufmann.

[25] Shan, C., Gong, S., McOwan, P.W., 2009. Facial expression recognition based on local binary patterns: A comprehensive study. Image and Vision Computing 27, 803–816.

[26] Tian, Y., 2004. Evaluation of face resolution for expression analysis, in: Computer Vision and Pattern Recognition Workshop.

[27] Valstar, M., Pantic, M., 2010. Induced disgust, happiness and surprise: an addition to the MMI facial expression database, in: International Language Resources and Evaluation Conference.

[28] Valstar, M., Patras, I., Pantic, M., 2005. Facial action unit detection using probabilistic actively learned support vector machines on tracked facial point data, in: IEEE Conference on Computer Vision and Pattern Recognition Workshop, pp. 76–84.

[29] Vapnik, V.N., 1995. The nature of statistical learning theory. New York: Springer-Verlag.

[30] Viola, P., Jones, M., 2001. Rapid object detection using a boosted cascade of simple features, in: IEEE Conference on Computer Vision and Pattern Recognition.

[31] Wallhoff, F., 2006. Facial expressions and emotion database. www.mmk.ei.tum.de/˜waf/fgnet/feedtum.html.

[32] Wang, W., Chen, W., Xu, D., 2011. Pyramid-based multi-scale lbp features for face recognition, in: International Conference on Multimedia and Signal Processing (CMSP), pp. 151 –155.

[33] Yang, P., Liu, Q., Metaxas, D.N., 2010. Exploring facial expressions with compositional features, in: IEEE Conference on Computer Vision and Pattern Recognition.

[34] Zhang, Y., Ji, Q., 2005. Active and dynamic information fusion for facial expression understanding from image sequences. IEEE Transactions on Pattern Analysis and Machine Intelligence 27, 699–714.

[35] Zhao, G., Pietikäinen, M., 2007. Dynamic texture recognition using local binary patterns with an application to facial expressions. IEEE Transaction on Pattern Analysis and Machine Intelligence 29, 915–928.

[36] Zhaoping, L., 2006. Theoretical understanding of the early visual processes by data compression and data selection. Network: computation in neural systems 17, 301–334.