Emotion recognition based on facial components

P ITHAYA RANI* and K MUNEESWARAN

Department of Computer Science and Engineering, Mepco Schlenk Engineering College, Mepco Nagar, Amathur (Post), Sivakasi 626 005, India
e-mail: muhilarani@gmail.com; kmuni@mepcoeng.ac.in

MS received 12 November 2015; revised 13 August 2016; accepted 26 December 2017; published online 28 March 2018

Abstract. Machine analysis of facial emotion recognition is a challenging and an innovative research topic in human–computer interaction. Though a face displays different facial expressions, which can be immediately recognized by human eyes, it is very hard for a computer to extract and use the information content from these expressions. This paper proposes an approach for emotion recognition based on facial components. The local features are extracted in each frame using Gabor wavelets with selected scales and orientations. These features are passed on to an ensemble classifier for detecting the location of face region. From the signature of each pixel on the face, the eye and the mouth regions are detected using the ensemble classifier. The eye and the mouth features are extracted using normalized semi-local binary patterns. The multiclass Adaboost algorithm is used to select and classify these discriminative features for recognizing the emotion of the face. The developed methods are deployed on the RML, CK and CMU-MIT databases, and they exhibit significant performance improvement owing to their novel features when compared with the existing techniques.

Keywords. Semi-local binary patterns; Gabor wavelet; multiclass Adaboost; normalization; face detection.

1. Introduction

One of the most interesting and challenging problems faced in the field of image analysis is the facial emotion recognition. It finds its major application in human–computer interaction. For the computer to interact with the human, it should possess the communication skills similar to those of human beings. One such skill is the ability to understand the emotional state of the person [1]. Research in the field of psychology and neurology has shown that facial expression is the most natural and primary cue for communicating the quality and nature of emotions, and it correlates well with the body and voice [2]. Frequently, all humans express their emotions with their facial expressions. Important facial emotions are accepted on the set of measurements, such as happiness, sadness, anger, fear, surprise and disgust as suggested [3]. The facial expression recognition plays a vital role in various applications in day-to-day developments and in other areas such as interactive video, virtual reality, video-conferencing, user profiling, games, intelligent automobile systems, entertainment industries and retrieval of image or video.

Techniques and issues for recognition of facial emotion have been examined extensively in the previous decade. A deformable 3D facial expression model for emotion recognition from video sequences is presented [4]. This

model was driven by the key fiducial points, and thus makes it possible to generate the intrinsic geometries of the emotional space. However, some of the features selected in a global scenario were redundant. A method based on texture feature is presented [5] for facial expression recognition, which combines image ratio features with facial animation parameters for describing the geometric motions of facial feature points. The work [6] stated that the facial features like eyebrows, eyes, mouth and nose are extracted using the SUSAN edge detection operator, facial geometry and edge projection analysis for face emotion recognition.

Ji and Idrissi [7] used a predesigned mask of 10 blocks on the face image. Then the local binary pattern (LBP) operator is applied on these 10 blocks, which concatenates the histograms related to the 10 blocks ($256 \times 10 = 2560$) for face expression recognition which is resource-intensive in terms of memory and time. Shan *et al* [8] proposed 16,640 LBP features extracted from the face images. The top 50 features are selected by Adaboost for each expression, which is mainly distributed in the eye and mouth regions. Ahmed Khana *et al* [9] reported eye and mouth features extracted using pyramid-based spatial representation of LBP (PLBP) features for emotion classification but it took significant memory and time. Liu *et al* [10] used a Boosted Deep Belief Network (BDBN) for performing the three training stages of feature learning/strengthening, feature selection and classifier construction in a unified loop

*For correspondence

framework. A two-stage Multitask Sparse Learning (MTSL) framework was proposed [11], which divided the facial image into local patches, and later LBP features were used to extract the local appearance of each patch. However, high classification rate was not achieved.

One of the most frequently used descriptors is the Gabor wavelet (GW) [12], which is powerful, but intensive both in time and memory, to convolve face images with a bank of Gabor filters to extract multi-scale and multi-orientation coefficients for recognizing the emotion. Praseeda Lekshmi *et al* [13] explain face expression classification from video sequences, where skin colour detection method is applied to detect face regions recognized using the Principal Component Analysis (PCA) method. Yi and Khalid [14] introduced two new textons containing Vertical Time Backward (VTB) and moments on spatiotemporal plane to describe the transformation of human face during facial expressions. We identified that in these methods, the specific regions of the face images such as the eyes, eyebrows and the mouth are mainly involved in recognizing the facial expression. In spite of many methods that have been explored in the literature for the facial expression recognition, still there is a lot of scope for the exploration of new algorithms to improve the performance. Our proposed algorithm addresses the exploration of GW feature for detection of face components and normalized semi-local binary pattern (NSLBP) features for emotion recognition. The advantage of proposed algorithm is that it delivers high recognition rate and is robust to variations in illumination. The rest of the paper is organized as follows. Section 2 describes the proposed work. Section 3 shows the experimental set-up and results. Section 4 offers conclusion and directions for future work.

2. The proposed work

The proposed work describes emotion recognition based on facial compounds. For finding the location of face components, initially there is a need to estimate the portion and the scale of the face region in each frame. For this purpose, local features are extracted in a given frame using GW [15]. These Gabor features are most suitable for face/non-face classification. These features are passed on to a boost classifier with well-trained signature of face and non-face images for detecting location of the face region in the frame. From the signature of each pixel in the face region, the eye and the mouth regions are detected using an ensemble classifier with well-known training images consisting of eye, non-eye face and mouth and non-mouth face signature of the images. A schematic diagram for detecting the face and its components is shown in figure 1.

For facial emotion recognition, eye and mouth features are extracted using NSLBP features. The multiclass AdaBoost algorithm is used to learn and classify these discriminative features such as happiness, sadness, anger, surprise, disgust, fear and neutral expressions.

2.1 Detection of face region

Initially, each frame is extracted and represented as a set of frames $\{f_1, f_2, \dots, f_n\}$ from video V , where ' n ' is the number of frames. For finding the location of face region in the k th frame f_k , GW features and a boost classifier are used, which is explained here.

2.1a GW: The Gabor features offer the best simultaneous localization of spatial and frequency information. It has been widely applied in image processing tasks such as edge detection, invariant object recognition and compression [15]. GW is a band pass filter and has been widely used in many applications, especially capturing the texture information, which when combined with other feature extraction techniques like moments and spectral distribution exhibits very good detection rate [16]. The 2D Gabor filter is adopted and it can be mathematically expressed as

$$\psi(a, b, \sigma, \theta) = \exp\left(-\frac{(A^2 + \gamma^2 B^2)}{2\sigma^2}\right) \cos\left(\frac{2\pi}{\lambda} A\right) \quad (1)$$

$$A = a \cos \theta + b \sin \theta, \quad B = -a \sin \theta + b \cos \theta$$

where orientation is θ , the effective width is σ , the wavelength is λ and the aspect ratio is γ . Typically, the f_k th frame, its convolution with a Gabor filter Ψ is defined as follows:

$$GF_{k,s,o}(a, b) = f_k(a, b) \Theta \Psi(a, b, s, o) \quad (2)$$

where Θ is the symbol for convolution. Eight scales $s \in (5-19$ with increments of 2) and four orientations $o \in (-45^\circ, 90^\circ, 45^\circ, 0^\circ)$ are used in the design of Gabor filters. The f_k th frame is convoluted with impulse of the Gabor filters as in Eq. (2), resulting in a series of Gabor images with features such as bars and edges usefully emphasized to facilitate the location of the face region in a better way. Accordingly, each pixel within a frame is delineated as $(\sigma \times \theta)$ 32-dimensional features.

2.1b Classification: The boost algorithm as proposed [17] for object detection exhibits a low false positive rate. The benefit of boost algorithm is to reduce the redundancies of high-dimensional feature space and computational expenses. It is sensitive to noisy data and outliers. Boost classifiers are constructed from a combination of the existing weak classifiers. Some of the boost classifiers such as RealAdaboost (RA) [18], GentleAdaboost [19] and ModestAdaboost (MA) [20] are variants of the boost algorithms, which are deployed in the proposed work to detect the faces in videos. Initially the training set is formed in Database-1 (DB₁) using the GW signatures from the collection of training images having both face and non-face images. The detection of human face in video sequences is described in algorithm 1 as follows.

Figure 1. Schematic diagram for detecting the face and face components.

Algorithm 1 for detection of face region in video sequences

Input : Video $V \leftarrow \{f_1 f_2, \dots, f_n\}$ set of frames where n is the number of frames. Set scale $\sigma = 8$ and orientations $\theta = 4$, $DB_1 \leftarrow$ GW signature of training images (face and non-face images).

Output : F-face region

Iterate : $k = 1$ – n in V // frame extraction;

$f_k(a, b) \leftarrow$ resize into 240×240 pixels in each frame f_k .

Iterate : $s = 1$ – σ , $o = 1$ – θ

$GF_{k,s,o}(a, b) \leftarrow$ extract GW feature $(f_k(a, b) \Theta \psi(a, b, s, o))$

End iterate // pixel-wise classification

$F_{k,s,o}(a, b) \leftarrow$ obtain the face region using boost classifier $(GF_{k,s,o}(a, b), DB_1)$

End iterate

2.2 Detection of facial component

The location and scale of eye and mouth regions are to be detected from the face region. Thus, the GW signatures corresponding to the database-2 (DB_2) with eye and non-eye face images and database-3 (DB_3) with mouth and non-mouth face images are created as the training set. A

similar face detection procedure is carried out for the extraction of eye and mouth regions from the pixel-wise signature of the face region. The result for eye region is represented as E and mouth region is represented as M. The detection of facial components is described in Algorithm 2.

Algorithm 2 for detection of eye and mouth region in video sequences

Input : Video $V \leftarrow \{f_1, f_2, \dots, f_n\}$ set of frames where n is the number of frames. $DB_2 \leftarrow GW$ signature of training images (eye and non-eye face images). $DB_3 \leftarrow GW$ signature of training images (mouth and non-mouth face images).

Output : E-eye region and M-mouth region

Iterate : $k = 1 \rightarrow n$ in V // frame extraction;
 $f_k(a, b) \leftarrow$ resize into 240×240 pixels in each frame f_k ;
 $F_{k,s,o}(a, b) \leftarrow$ obtain the signature of face region from each frame $f_k(a, b)$ using algorithm 1.
 $FS_{k,s,o}(a, b) \leftarrow$ resize the obtained signature of face region $F_{k,s,o}(a, b)$ into 110×110 .
 $E_k(a, b) \leftarrow$ Obtain the eye region from pixel-wise GW signature of face region using the boost classifier $(FS_{k,s,o}(a, b), DB_2)$
//classify into eye and non-eye face regions.
 $M_k(a, b) \leftarrow$ obtain the mouth region from pixel-wise GW signature of face region using the boost classifier $(FS_{k,s,o}(a, b), DB_3)$
//classify into mouth and non-mouth face regions

2.3 Facial emotion recognition

The existing LBP [8] method labels the pixels of an image by thresholding a 3×3 neighbourhood of each pixel with the centre value and considering the results as a binary number (see figure 2 for an illustration). Then the histogram of the labels can be used as a texture descriptor. This method achieved only average performance. Moreover, the LBP is sensitive to noise because the point features compare a central pixel with neighbouring pixels. To overcome this drawback, Kyungjoong *et al* [21] proposed a new feature named semi-LBP (SLBP), a set of binary patterns with embedded local regions. A procedure in this SLBP system, the sum of sub-regions, reduces the influences of noise. In addition, it encodes the comparative sizes of the central region with locally neighbouring regions into a binary code as in an LBP feature (see figure 2). The noisy and low-quality image frames can be overcome using SLBP features. The intensity variation due to different factors affects the recognition rate. Hence, to make intensity values invariant, a normalization (N) technique is proposed, which is combined with SLBP features (NSLBP). This feature is simple, has high discriminative power and also computationally efficient. The advantage of NSLBP is that it is robust to illumination variations and insensitive to noisy images. The eye and the mouth features are extracted using NSLBP and the facial expression is classified using a multi-class Adaboost classifier, which is explained here.

2.3a NSLBP-based feature extraction: The locations of the eye and the mouth regions are used for facial emotion recognition. The grey levels of the eye region E_k in the k th frame is normalized as follows:

$$En_k = N(E_k) = \sum_{i=0}^{L-1} \frac{p_i}{T}, \quad i = 0, 1, \dots, L-1 \quad (3)$$

where En_k represents the normalized eye regions, N denotes the normalization function, T is the total number of pixels, p_i is the number of pixels with grey level E_i and L is the number of discrete grey levels. Following the intensity normalization, the normalized eye region in the k th frame is divided equally into $2 \times 2 = 4$ blocks (non-overlapping blocks), which is represented as B ($B = 1, 2, \dots, (2 \times 2)$). Typically, each block is of the size of 13×40 pixels as follows:

$$En_{k,B}(a, b) = En_k(a : a + 12, b : b + 39) \quad (4)$$

where $a = 1 : r_1$ increment by 13 ;
 $b = 1 : c_1$ increment by 40

where r_1 and c_1 are the number of rows and columns in the eye region, respectively. Each block is again divided into overlapping sub-blocks. It is represented as SB ($SB = 1, 2, \dots, (m_1 \times n_1)$) where each sub-block is of size 4×4

Figure 2. LBP vs. SLBP vs. NSLBP procedure.

pixels. The number of sub-blocks is $(m_1) \times (n_1)$ where m_1 and n_1 are the number of rows and columns in each block, respectively. The NSLBP is defined by a binary coding function to find the sum of intensity in a window of size 2×2 pixels within a sub-block as follows:

$$S_{k,B}^{p,q,(a_1,b_1)} \varepsilon_{SB}(a,b) = \sum_{a_2=u}^{a_2+1} \sum_{b_2=v}^{b_2+1} En_{k,B}(a+a_1+a_2, b+b_1+b_2)$$

(5)

where $1 \leq p \leq m_1$, $1 \leq q \leq n_1$, $a_1 = p : p + 3$, $b_1 = q : q + 3$, $1 \leq u \leq 3$, $1 \leq v \leq 3$

where $S_{k,B}^{p,q,(a_1,b_1)} \varepsilon_{SB}(a,b)$ is the sum of intensities in each sub-block 'SB', whose size is 4×4 at each pixel position (a,b) . The resultant value is available as 3×3 matrixes for each sub-block at pixel position (a,b) . The centre value of 3×3 matrixes is compared to eight neighbour values. An 8-bit code is coined, which represents the value in the decimal form at each pixel position (a,b) in the block. The representations of these quantities are as follows.

Let M represent the matrix defined as follows:

$$M = S_{k,B,(a_1,b_1) \in SB}^{p,q}(a,b). \quad (6)$$

The value using the proposed method NSLBP is obtained as

$$NSLBP(i_1, j_1) = \sum_{n_1=0}^7 T(i_1, j_1) 2^{n_1} \quad (7)$$

where $0 \leq i_1 \leq 2$, $0 \leq j_1 \leq 2$, $i_1 \neq 1$ and $j_1 \neq 1$ and

$$T(i_1, j_1) = \begin{cases} 1 & M(i_1, j_1) \geq M(1, 1), \\ 0 & \text{else.} \end{cases} \quad (8)$$

The NSLBP method $NSLBP_{P,R}$ produces 2^P different output values, corresponding to the 2^P different binary patterns that can be formed by the P pixels in the neighbour set. The notation (P,R) denotes a neighbourhood of P equally spaced sampling points on a circle of radius R . Therefore, it is possible to use only a subset of the 2^P LBPs to describe the texture of images. Ojala *et al* [22] called these fundamental patterns as uniform patterns. The NSLBP is called uniform if it contains at most two bitwise transitions from 0 to 1 or vice versa when the binary string is considered circular. For instance, 11111111 and 11100011 are uniform patterns. It is observed that most of the texture information is contained in the uniform patterns. Accumulating the patterns that have more than 2 transitions into a single bin yields an NSLBP operator, denoted as $NSLBP_{P,R}^{u2}$, with less than 2^P bins. For example, the number of labels for a neighbourhood of 8 pixels is 256 for the standard NSLBP but 59 for $RLBP^{u2}$ (superscript u2 stands for uniform patterns). After obtaining the value using NSLBP method for each pixel associated with a block in the eye region, a histogram of the labelled image can be used as a texture descriptor. Following this setting [22], all values are quantified into 59 bins ($NSLBP_{8,1}^{u2}$) according to the uniform strategy. It gives a good trade-off between recognition performance and feature vector length. The histograms of each block in the eye region $NSLBPe$ can be defined as follows:

$$NSLBPe_L = \sum_{p,q} I(NSLBP(p,q) = L) \quad (9)$$

where $1 \leq p \leq m_1$, $1 \leq q \leq n_1$, $a_1 = p : p + 3$, $b_1 = q : q + 3$, $1 \leq L \leq 59$ where L is the number of bins for the values produced by the NSLBP operator.

$$I(A) = \begin{cases} 1 & A \text{ is true,} \\ 0 & A \text{ is false.} \end{cases} \quad (10)$$

Concatenate the $NSLBPe_L$ histograms that are extracted from all blocks in the eye region into a single feature. As a result we obtained $(2 \times 2 \times L)$ 236 NSLBP histogram bins in the eye region. Similar to the eye region, the NSLBP feature extraction method is carried out for the mouth region. It is represented as ($NSLBPm_L$). These two histograms are concatenated into a single spatially enhanced feature of histograms as follows:

$$NSLBPe_m = [NSLBPe_L \cup NSLBPm_L]. \quad (11)$$

The NSLBP histograms contain information about the description of the local micro-patterns such as edges, spots and flat areas over the whole image. Hence, it is used to statistically describe image characteristics. As a result we obtained 472 $NSLBPh$ histogram bins in the k th frame.

2.3b Learning discriminative NSLBP histograms: During the feature extraction stage, each training image contains 472 NSLBP histograms. To reduce the feature dimension and the redundant information, it is necessary to select the discriminative NSLBP features. We propose to learn discriminative NSLBP features for better facial expression recognition. Multiclass Adaboost provides a simple yet effective approach for stage-wise learning of a non-linear classification function [23, 24]. Hence, the multi-class AdaBoost algorithm is adopted to learn the discriminative NSLBP histograms with the help of training images. It is not only useful to select the discriminating features but also to train the classifier. The multiclass AdaBoost learns a small number of weak classifiers whose performance is just better than random guessing, and boosts them iteratively into a strong classifier of higher accuracy/minimum error. The process of multiclass AdaBoost maintains a distribution of the training samples. At each iteration, a weak classifier that minimizes the weighted error rate is selected, and the distribution is updated to increase the weights of the misclassified samples to reduce the importance of the others. Similar to the proposal [17], the weak classifier consists of a feature that corresponds to a single NSLBP histogram bin.

The proposed work for recognizing the facial expression is summarized in Algorithm 3. Initially, the training set is formed in Database-4 (DB₄) using the NSLBP signatures from the collection of training images having happiness, sadness, fear, anger, surprise, disgust and neutral expression images. The recognition performance of the boosted strong classifier is achieved as a function of the number of discriminative features selected from the DB₄ and it is represented as $\check{T}$.

Algorithm 3 for recognizing the facial emotion in video sequences

Input : Label the f_k th frame in video V (example $k = 1$). Set scale $\sigma = 8$ and orientation $\theta = 4$. Set histogram bins $L = 59$. The training databases are DB₁, DB₂, DB₃ and DB₄. $\tilde{T}$ – selected discriminative NSLBP signature from DB₄ using the multiclass Adaboost algorithm.

Output : F – facial emotion (happiness, sadness, fear, disgust, anger, surprise and neutral);

Start : $f_k(a, b) \leftarrow$ resize into 240×240 pixels in given frame f_k .

// Obtain the eye and mouth regions using Algorithm 2

// **Extract NSLBP features**

$En_k(a, b) \leftarrow$ Apply normalization techniques in the eye region $E_k(a, b)$.

$Mn_k(a, b) \leftarrow$ Apply normalization techniques in the mouth region $M_k(a, b)$

where r_1 and c_1 are the number of rows and columns in the eye region, respectively;

$NSLBP_{e_L} = []$ // initialize the empty feature for eye region.

$B = 1$ // initialize the block

Iterate 1: $a = 1:13:r_1$; $b = 1:40:c_1$ // normalized eye region is divided into $2 \times 2 = 4$ blocks

$En_{k,B}(a, b) = En_k(a + 12, b + 39)$

// m_1 and n_1 are the number of rows and columns of each block ($En_{k,B}$), respectively.

Iterate 2: $p = 1:m_1, q = 1:n_1$

// Each block is divided into overlapping sub-blocks (SBs), where an 'SB' contains 4×4 pixels

and sum of windows each of size 2×2 pixels used within an 'SB' as follows:

$$S^{p,q}_{k,B,(a_1,b_1) \in SB}(a, b) = \sum_{a_2=u}^{a_1+1} \sum_{b_2=v}^{b_1+1} En_{k,B}(a + a_1 + a_2, b + b_1 + b_2)$$

where $a_1 = p : p + 3, b_1 = q : q + 3, 1 \leq u \leq 3, 1 \leq v \leq 3$

// The resultant value is reduced to 3×3 pixels from 4×4 blocks. Then the LBP is obtained from

// resultant 3×3 pixels. It returns one scalar value, which is stored at pixel position (p, q) . The

// representations of these quantities are as follows.

Let M represent the matrix defined as follows:

$$M = S^{p,q}_{k,B,(a_1,b_1) \in SB}(a, b).$$

The value using the proposed method NSLBP is obtained as follows:

$$NSLBP(i_1, j_1) = \sum_{n_1=0}^7 T(i_1, j_1) 2^{n_1} \quad 0 \leq i_1 \leq 2; 0 \leq j_1 \leq 2 \text{ where } i_1 \neq 1 \text{ and } j_1 \neq 1.$$

$$T(i_1, j_1) = \begin{cases} 1 & M(i_1, j_1) \geq M(1, 1), \\ 0 & \text{else.} \end{cases}$$

end iterate 2

$NSLBP_L \leftarrow$ apply histograms $\{NSLBP(p, q), L\}$

$NSLBP_{e_L} \leftarrow [NSLBP_{e_L} \cup NSLBP_L]$ // NSLBP histograms for eye region

$B \leftarrow B + 1$ // move to next block

end iterate 1

// Similar to the eye region, the NSLBP features extraction procedure is carried out for the mouth region ($NSLBP_{m_L}$). These two histograms are concatenated into a single spatially enhanced feature of histograms as follows:

$$NSLBP_{em} = [NSLBP_{e_L} \cup NSLBP_{m_L}]$$

$F \leftarrow$ obtained facial expression using Multiclass Adaboost classifier ($NSLBP_{em}, \tilde{T}$)

Stop

3. Experimental results and discussion

To evaluate the performance of the proposed method, RML facial expression video data set [25], Cohn–Kanade (CK) database [26] and CMU-MIT database [27] were used for the experiment. The RML video data set was originally developed for language- and context-independent emotional recognition. It contains 720 video face sequences of 8 subjects with 1–6 emotions per subject in a nearly frontal view with the position of the head not stable, including a simple background. Each video picture starts with a single emotional expression and ends at the apex of that expression, while the first frame of every video sequence shows a neutral face. Totally, 720 video sequences are selected, which include 119 of happiness, 120 of sadness, 122 of fear, 120 of surprise, 120 of disgust and 119 of anger expressions.

CK is one of the most popular databases in current face image analysis community. The database consists of 100 university students aged from 18 to 30 years, of which 65% are female, 15% are African-American and 3% are Asian or Latino. In our experiments, 348 sequences are selected from the CK database, which includes 40 of anger, 41 of disgust, 45 of fear, 97 of happiness, 48 of sadness and 77 of surprise data. Each sequence runs from a face in neutral state to a target expression in peak state. All images from this database show faces with frontal view or nearing frontal view, and with the position of the heads as stable. The CMU-MIT face data set is one of the early benchmarks for face detection. The data set contains 130 grey-scale images with a total of 511 frontal view faces, some of which are noisy images. This database is not used for emotion recognition as the person does not show the seven expressions. For computational simplicity, in both databases, the resolutions were scaled down to 240×240 pixels. Table 1 shows the detailed contents of the training and testing data sets for face and its component detection. Figure 3 shows some samples of the training images.

3.1 Result for detection of face region

3.1a Performance of GW features: In our experiments, the orientation and scale of Gabor filters imposed on images are two key parameters that determine the effectiveness of the extracted texture features. Table 2 compares the detection results obtained using 2, 4, 6 and 8 orientations of Gabor filters (the number of the scales is fixed at 8) using a GA classifier and RML database. The four sets of orientations are $(90^\circ, 0^\circ)$, $(-45^\circ, 90^\circ, 45^\circ, 0^\circ)$, $(-45^\circ, -22.5^\circ, 0^\circ, 22.5^\circ, 45^\circ, 90^\circ)$ and $(90^\circ, 67.5^\circ, 45^\circ, 22.5^\circ, 0^\circ, -22.5^\circ, -45^\circ, -67.5^\circ)$, correspondingly. An increase from 4 to 6 and 8 orientations lowers the performance by

2% and 3%, respectively. However, 8 orientations have better performances of 2% than 2 orientations performance at all resolutions. In addition, a decrease of the training image resolutions from 110×110 to 90×90 lowers the performance by 3% in all different orientations. However, the 90×90 resolution training images have higher performance of 2% than 125×125 resolution training images. It can be seen that utilizing the default 4 orientations has the highest accuracy of 98% among the four sets of orientation values in 110×110 resolutions of training images. This result suggests that Gabor filters require at least 4 orientations to capture most of the discrimination information.

Table 3 compares the accuracies obtained using 4, 6, 8 and 10 scales of Gabor filters (the number of orientations of Gabor filters is fixed at four) in each resolution training image. The four sets of scales are composed of $(5:2:11)$, $(5:2:15)$, $(5:2:19)$ and $(5:2:23)$ pixels, correspondingly. A decrease from 8 to 6 and 4 scales reduces performance by 2% and 4%, respectively. However, the 10 scales have higher performance of 1% than 6 scales performance at all resolutions of training images. Moreover, the 110×110 resolutions of the training images have higher performance of 2% than 125×125 resolutions of the training images in all orientations. The 125×125 resolution training images have a slight increase of 1% than 90×90 resolution training images. Similar to the comparison results for different orientations of Gabor filters in table 2, using 8 scales of Gabor filters delivers the highest accuracy of 98% among the four sets of scale values. In both tables 2 and 3, 110×110 resolutions training images show the best result. Therefore, the training image resolution 110×110 is a suitable size for detection of face and its components. The results also confirm that the default 4 orientations and 8 scales of Gabor filters are the optimal parameters for detecting the face and its components such as eye and mouth.

3.1b Classifier performance: The GW features are trained through variant Adaboost classifiers such as RA, GA and MA in 110×110 resolutions training images. They are compared for error checking with 50 boosting iterations as shown in figure 4. The analysis of the GA returns error rate of 0.024, which is less than 0.055 and 0.063 error rates of RA and MA, respectively. Hence, GA is selected as the detection algorithm for our system.

Due to space limitation, we display only the face detection (sensitivity) rate of 98% using RML database. We obtained the face detection rate of 98.5% and 99% using CK and CMU-MIT databases, respectively. Figure 5 shows the sample face detection result in both the databases. The size of the bounding box is determined using the scale on the detected face on the video sequences. From figure 5, we observe that the proposed GW features are robust against the different expression variations.

Table 1. Description of training and test data selected from the RML, CK and CMU-MIT databases.

Collected no of training images in each database					Feature extraction in each image [GW: 4 scales, 8 orientations ($4 \times 8 = 32$)]	Using three types of classification	Testing data set
Collected from RML and CK databases separately	Three types of resolution in each image	Manually created	Resolution in each image	Total no. of training images			
350 faces (different poses, expressions)	a) 90×90 b) 110×110 c) 125×125	350 non-face images	a) 90×90 b) 110×110 c) 125×125	700	a) $90 \times 90 \times 32$ b) $110 \times 110 \times 32$ c) $125 \times 125 \times 32$	RA GA MA	720 videos— RML 348 face sequences— CK CMU-MIT face detection database
150 eye images	a) 35×100 b) 20×65 c) 26×80	150 non-eye face images	a) 110×110	300	$35 \times 100 \times 32$ $20 \times 65 \times 32$ $26 \times 80 \times 32$ $110 \times 110 \times 32$		
150 mouth images	a) 35×65 b) 20×45 c) 26×60	150 non-mouth face images	a) 110×110	300	$35 \times 65 \times 32$ $20 \times 45 \times 32$ $26 \times 60 \times 32$ $110 \times 110 \times 32$		

Figure 3. Sample of training images.**Table 2.** Detection performances (%) obtained using four sets of orientation.

Orientations	Training image resolution		
	90×90	110×110	125×125
2	90	93	88
4	95	98	93
6	93	96	91
8	92	95	90

Table 3. Detection performances (%) obtained using four sets of scales.

Scale	Training image resolution		
	90×90	110×110	125×125
4	91	94	92
6	93	96	94
8	95	98	96
10	94	97	95

Figure 4. Error rates from three Adaboost algorithms in RML.

3.2 Results for detection of facial components

For the training, the eye and non-eye and mouth and non-mouth face images are made available. Detection of the facial components is done for the eye and mouth in a

sequential fashion with different resolutions of $(35 \times 100, 20 \times 65$ and $26 \times 80)$ for eye and $(35 \times 65, 20 \times 45$ and $26 \times 60)$ for the mouth. In all these experiments, pixel-based GW features are used. The results are shown in figure 6a and b, which represents the accuracy of the component detection with different resolutions in different classifiers such as RA, MA and GA. The high accuracy (sensitivity) rate is obtained as 92% in RML, 91% in CK in the resolution 26×80 for eye and as 93% in RML, 92% in CK in the resolution 26×60 with GA classifier for mouth. These eye and mouth resolutions are used for facial expression recognition. The results for facial compound detection are shown in figure 7. The size of the bounding box is determined using the scale of the detected facial components on its face images.

3.3 Performance comparison with the previous approach for detection of face, eye and mouth

The proposed work is carried out using a PC with Intel Core i5 @ 3.20 GHz, 8 GB RAM specification and the tool used was Matlab 2013a version for all the test cases. For comparison purpose, we have selected the existing Haar feature, which is used by Viola and Jones [17] for face detection. Though it is old, it is a well-known method. Hence many

Figure 5. Example face detection result in different databases: (a) RML, (b) CK and (c) CMU-MIT.

Figure 6. Result for face component detection: (a) RML and (b) CK.

Figure 7. Sample result of face component detection for different emotional faces.

authors [8, 5, 9, 11] have recently used it for face detection. In their approach [17], the method is divided into three steps. The first step is integration of image representation called integral image. The integral image can be computed from an image by summing up the pixels that are above and to the left of (x, y) . Once computed, three kinds of Haar features (two-rectangle, three-rectangle and four-rectangle feature) are extracted. By means of this, totally 160,000 rectangle features are extracted. The second step is selecting the salient feature from the 160,000 rectangle features using the AdaBoost classifier algorithm. The third step is

combining the classifier in a cascade structure for detecting the face region. Ahmed Khana *et al* [9] adopted the Viola and Jones [17] method for detecting the face, eye and mouth regions. Thus, we have measured and compared the proposed GW features with the Haar features. The results of the detection rate and time for detecting the face, eye and mouth regions in both the databases are shown in tables 4 and 5, respectively. We infer from both the tables that the proposed GW features achieved better detection rate and execution time is much lower than the Haar features in both the databases.

3.4 Results for facial emotion recognition

The facial emotions are recognized from video sequences and are formulated as a sequence labelling problem by labelling every frame with the correct facial emotion and neutral state. Initially in our experiment, 720 video sequences are selected from the RML emotion database. From each video sequence, a neutral face frame and two peak frames are labelled as proposed [9, 10] for prototypic emotion recognition, resulting in 1640 frames, which include 238 of happiness, 240 of sadness, 244 of fear, 240 of surprise, 240 of disgust, 238 of anger and 200 of neutral frames. Next, 348 sequences are selected from the CK database. From each sequence, the neutral and three peak frames are used for expression recognition, resulting in 1300 images, which include 291 of happiness, 144 of sadness, 135 of fear, 231 of surprise, 123 of disgust, 120 of anger and 256 of neutral images. For the evaluation of our work, we adopted a 10-fold cross-validation testing scheme in RML and CK databases.

3.4a Performance of discriminative feature selection: We plot in figure 8 the minimum error performance of the boosted strong classifier as a function of the number of features selected in both databases. With the selected 27 histogram bins, the boosted strong classifier achieved the minimum error rate in RML and CK. From figure 8a, it can be observed that the NSLBP histograms have an error rate of 0.034, which is less than the SLBP features' error rate of 0.06 in RML data set. From figure 8b, it can be observed that the NSLBP histograms have an error rate of 0.0345, which is less than the SLBP features' error rate of 0.054 in CK data set. The proposed NSLBP features reduced the minimum error.

3.4b Performance of NSLBP features: Tables 6 and 7 compare the accuracies obtained using LBP ([8], PLBP [9], SMCT, SLBP [21] and NSLBP (**Proposed**) histograms examined in RML and CK databases, respectively. In tables 4 and 5, the LBP and PLBP features have lower performance of 3% in RML, 4% in CK and 2% in RML, 2% in CK, respectively, than SLBP features performance. Both databases contain people with different skin colours of the faces. To enhance the expression recognition process, it is necessary to bring the skin condition to a normal form.

This suggests that there is scope for further improvement by including the N techniques in SLBP features. The average facial expression recognition rate has considerably improved to about 4% and 2% by incorporating the N techniques in RML and CK databases, respectively. The recognition rate of the CK database is slightly lower than that of the RML database since the CK database has the person's expression such as sad and neutral as the same. As a result, we conclude that the NSLBP features result in better facial emotion recognition. Moreover, tables 8a, b and 9a, b show the proposed SLBP features and NSLBP features upgraded in both the databases, respectively.

3.5 Performance comparison with the previous approach

In terms of time and memory costs of feature extraction process in recognizing the emotion, we have measured and compared our descriptor with the LBP [8] and PLBP [9] features, which is shown in table 10. It can be seen that the proposed NSLBP descriptor's memory consumption and its execution time are much lower than the other compared features.

A comparison of the achieved recognition rate of the proposed algorithm with the reported result in state-of-the-art methods is shown in table 11. These reported results are selected because they use a similar testing strategy and same databases. When seven emotions are used, the recognition result of the proposed approach is 5.8%, 5.1%, 6.3% and 8.8% higher than those used LBP [7], LBP [8], boosted LBP [8] and geometric features [5], respectively, using CK database. The total number of boosted LBP [8] features selected is 50, whereas in the proposed work, 27 boosted NSLBP features are selected for each expression. However, the proposed work's recognition rate achieved is higher than that by the boosted LBP features. The NSLBP feature recognition rate is 1.6% higher than those extracted by the eye and the mouth features [9] using CK database. It shows that the proposed NSLBP features have achieved the best recognition rate among the several related works. Moreover, the proposed work has demonstrated its ability to reduce the memory and time consumption.

Table 4. Comparison of detection accuracy with existing approach for detecting face, eye and mouth where R, C and C-M denote the RML, CK and CMU-MIT databases, respectively.

Feature	Recall (%)							Precision (%)						
	Face			Eye		Mouth		Face			Eye		Mouth	
	R	C	C-M	R	C	R	C	R	C	C-M	R	C	R	C
Haar	94	95	97.5	88	89	90	90	91	92	96	88	89	88	88
GW	98	98.5	99	92	91	93	92	94	95	97	91	90	90	90

Table 5. Comparison of time consumption with existing approach for detecting face, eye and mouth.

Database	Feature	Execution time of features extraction (s)			Execution time for detection in a single frame (s)		
		Face	Eye	Mouth	Face	Eye	Mouth
RML	Haar	0.030	0.021	0.016	0.0088	0.018	0.016
	GW	0.027	0.018	0.014	0.0078	0.015	0.013
CK	Haar	0.028	0.020	0.015	0.0084	0.016	0.015
	GW	0.025	0.018	0.013	0.0068	0.010	0.009
CMU-MIT	Haar	0.029	—	—	0.0083	—	—
	GW	0.025	—	—	0.0075	—	—

(a)

(b)

Figure 8. Discriminative features selected: (a) RML and (b) CK.**Table 6.** Result for facial emotion recognition rate in % (recall).

Different expressions	RML database					CK database				
	LBP	PLBP	SMCT	SLBP	NSLBP	LBP	PLBP	SMCT	SLBP	NSLBP
Happiness	91	93	94	97	99	90	94	95	97	99
Sadness	90	91	92	94	97	89	92	91	93	96
Fear	90	91	90	90	95	90	93	94	94	95
Surprise	92	96	94	97	99	91	92	94	96	98
Disgust	91	90	93	94	98	91	93	93	93	95
Anger	90	92	93	94	98	91	92	93	94	95
Neutral	90	89	92	95	98	91	92	92	95	97
Overall accuracy	91	92	93	94	98	91	93	93	95	97

Table 7. Result for facial emotion recognition rate in % (**precision**).

Different expressions	RML database (precision)					CK database (precision)				
	LBP	PLBP	SMCT	SLBP	NSLBP	LBP	PLBP	SMCT	SLBP	NSLBP
Happiness	90	89	92	93	96	91	93	93	94	97
Sadness	89	91	92	94	96	90	92	94	94	96
Fear	91	90	91	93	95	89	90	92	93	95
Surprise	92	92	93	94	97	90	91	93	94	97
Disgust	90	92	93	94	95	89	91	93	94	96
Anger	90	91	92	93	94	90	90	92	93	95
Neutral	89	89	91	93	93	91	92	92	94	96
Overall accuracy	90	91	92	93	95	90	91	93	94	96

Table 8. a. Confusion matrix using SLBP in RML. b. Confusion matrix using NSLBP in RML.

Desired	Happy	Sad	Detected fear	Surprise	Disgust	Anger	Neutral
<i>(a)</i>							
Happy	97	0	1.4	0	0	0	1.6
Sad	0	94	0	0	1	2	3
Fear	3	0	90	0	3	0	4
Surprise	1	0	2	97	0	0	0
Disgust	0	1.5	2.5	0	94	2	0
Anger	0	1	0	1	2	94	2
Neutral	2	0	0	0	1.5	1.5	95
<i>(b)</i>							
Happy	99	0	1	0	0	0	0
Sad	0	97	0	0	0	1	2
Fear	1.5	1.2	95	0	1	0	1.3
Surprise	0	0	1	99	0	0	0
Disgust	0	0	0	0	98	2	0
Anger	0	1	0	0	1	98	0
Neutral	1	0	0	0	0	1	98

4. Conclusion and future work

This paper investigated the benefits of emotion recognition based on facial compounds using NSLBP features. The experimental results show that the proposed method gives promising results when compared with the existing methods. The result is enhanced for the following reasons.

- For selecting only the facial compound information related to face emotion identity while discarding the

information that is not related to emotion identification.

- Features extracted using the proposed normalized SLBPs compensate for illumination variations and also have strong discriminative ability as the recognition result for seven facial expressions. It provides necessary information for classification.
- NSLBP features permitted high recognition rate in lesser time.
- The usage of memory also less.

Table 9. a. Confusion matrix using SLBP in CK. b. Confusion matrix using NSLBP in CK.

Desired	Happy	Sad	Detected fear	Surprise	Disgust	Anger	Neutral
(a)							
Happy	97	0	1.4	0	0	0	1.6
Sad	0	93	3	0	0	0	4
Fear	3	0	94	0	2	0	1
Surprise	2	0	2	96	0	0	0
Disgust	0	1.4	1.4	0	93	2.8	1.4
Anger	0	2	0	0	2	94	2
Neutral	2	0	0	0	2	1	95
(b)							
Happy	99	0	0	0	0	0	1
Sad	0	96	0	0	0	2	2
Fear	2	0	95	0	2	0	1
Surprise	1	0	1	98	0	0	0
Disgust	0	1	2	0	95	2	0
Anger	0	1	0	0	3	95	1
Neutral	1	0	0	0	1	1	97

Table 10. Comparison of memory and time consumptions.

Feature	Memory (feature dimension)	Execution time of features extraction (s)		Execution time for facial emotion recognition (s)	
		RML	CK	RML	CK
LBP	2478	0.05	0.03	0.20	0.18
PLBP	590	0.02	0.01	0.17	0.15
SMCT	412	0.01	0.01	0.15	0.15
SLBP	412	0.0085	.0061	0.12	0.1
Proposed NSLBP	412	0.01	.00676	0.14	0.13

A comparison of the achieved recognition rate of the proposed algorithm with the reported result in state-of-the-art methods is shown in table 11.

Table 11. Comparison with the state-of-the-art methods for emotion recognition.

Author	Database	Feature	Class	Fold	Recognition rate (%)
Yun and Ling [4]	RML	Geometric	7	–	88.2
Mingli <i>et al</i> [5]	Cohn–Kanade	Geometric	7	10	88.9
Ji and Idrissi [7]	Cohn–Kanade	LBP	7	10	91.9
Shan <i>et al</i> [8]	Cohn–Kanade	LBP	7	10	92.6
		Boosted LBP	6	10	95.1
		Boosted LBP	7	10	91.4
Ahmed Khana <i>et al</i> [9]	Cohn–Kanade	PLBP	6	10	96.7
Liu <i>et al</i> [10]	Cohn–Kanade	BDBN	6	10	96
Zhong <i>et al</i> [11]	Cohn–Kanade	MTSL	6	10	91
Proposed work	RML	NSLBP	6	10	98.5
		NSLBP	7	10	96.4
	Cohn–Kanade	NSLBP	6	10	98.3
		NSLBP	7	10	97.7

The test results show that NSLBP features find their appropriate role in facial expression recognition carried out on the RML and CK databases. The current work of

recognizing the facial expressions does not consider temporal information and occlusion, which will be addressed in future.

References

- [1] Fakhreddine K, Milad A, Jamil A S and Mo N A 2008 Human–computer interaction: overview on state of the art. *Int. J. Smart Sens. Intell. Syst.* 1(1): 23
- [2] Cohn J F 2010 Advances in behavioral science using automated facial image analysis and synthesis. *IEEE Signal Process.* 27(6): 128–133
- [3] Ekman P and Friesen A 1978 *The facial action coding system*. San Francisco, CA: Consulting Psychologist Press
- [4] Yun T and Ling G 2013 A deformable 3-D facial expression model for dynamic human emotional state recognition. *IEEE Trans. Circuits Syst. Video Technol.* 23(1): 142–157
- [5] Mingli S, Dacheng T, Zicheng L and Xuelong L 2010 Image ratio features for facial expression recognition application. *IEEE Trans. Syst. Man Cybernet. B: Cybernet.* 40(3): 779–788
- [6] Khandait S P, Thool R C and Khandaitm P D 2011 Automatic facial feature extraction and expression recognition based on neural network. *Int. J. Adv. Comput. Sci. Appl.* 2(1): 113–118
- [7] Ji Y and Idrissi K 2010 Learning from essential facial parts and local features for automatic facial expression recognition. In: *Proceedings of CBMI, 8th International Workshop on Content-Based Multimedia Indexing*
- [8] Shan C, Gong S and McOwan P 2009 Facial expression recognition based on local binary patterns: a comprehensive study. *Image Vis. Comput.* 27: 803–816
- [9] Ahmed Khana R, Alexandre B, Konika H and Bouakaza S 2013 Framework for reliable, real-time facial expression recognition for low resolution images. *Pattern Recogn. Lett.* 34: 1159–1168
- [10] Liu P, Han S and Meng Z 2014 Facial expression recognition via a boosted deep belief network. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1805–1812
- [11] Zhong L, Liu Q, Yang P, Huang J and Metaxas N 2015 Learning multiscale active facial patches for expression analysis. *IEEE Trans. Cybernet.* 45(8): 1499–1510
- [12] Littlewort G C, Bartlett M S and Lee K 2009 Automatic coding of facial expressions displayed during posed and genuine pain. *Image Vis. Comput.* 27: 1797–1803
- [13] Praseeda Lekshmi V, Sasikumar M and Naveen S 2008 Analysis of facial expressions from video images using PCA. In: *Proceedings of WCE*, London, UK
- [14] Yi J and Khalid I 2012 Automatic facial expression recognition based on spatiotemporal descriptors. *Pattern Recogn. Lett.* 33: 1373–1380
- [15] Lee T S 1996 Image representation using 2d Gabor wavelets. *IEEE Trans. PAMI* 18(10): 959–971
- [16] Muneeswaran K, Ganesan L, Arumugam S and Ruba Soundar K 2006 Texture image segmentation using combined features from spatial and spectral distribution. *Pattern Recogn. Lett.* 27: 755–764
- [17] Viola P and Jones M 2001 Rapid object detection using a boosted cascade of simple features. In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 1, pp. 511–518
- [18] Bo W, Haizhou A, Chang H and Shihong L 2004 Fast rotation invariant multi-view face detection based on RealAdaboost. In: *Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition*, pp. 79–84
- [19] Jerome F, Trevor H and Robert T 2000 Additive logistic regression: a statistical view of boosting. *Ann. Stat.* 28(2): 337–407
- [20] Vezhnevets A and Vezhnevets V 2005 Modest AdaBoost—teaching AdaBoost to generalize better. In: *Proceedings of Graphicon*, Akademgorodok, Novosibirsk
- [21] Kyungjoong J, Choi J and Jang G 2015 Semi-local structure patterns for robust face detection. In: *IEEE Signal Process. Lett.* 22(9): 1400–1403
- [22] Ojala T, Pietikainen M and Harwood D 1996 A comparative study of texture measures with classification based on features distributions. *Pattern Recogn.* 29(1): 51–59
- [23] Freund Y and Schapire R 1997 A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst. Sci.* 55: 119–139
- [24] Saberian M and Vasoncelos J 2011 Multiclass boosting: theory and algorithms. In: *Proceedings of Neural Information Processing Systems (NIPS)*, Granada, Spain, pp. 2124–2132
- [25] Yongjin W and Ling G 2008 Recognizing human emotional state from audiovisual signals. *IEEE Trans. Multimedia* 10(5): 659–668
- [26] Kanade T, Cohn J F and Yingli T 2000 Comprehensive database for facial expression analysis. In: *Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition (FG)*
- [27] Rowley H, Baluja S and Kanade T 1998 Rotation invariant neural network-based face detection. In: *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*