Multimed Tools Appl (2017) 76:7803–7821
DOI 10.1007/s11042-016-3418-y

Facial expression recognition based on local region specific features and support vector machines

Deepak Ghimire[1] · **Sunghwan Jeong**[1] · **Joonwhoan Lee**[2] · **San Hyun Park**[1]

Received: 19 August 2015 / Revised: 10 January 2016 / Accepted: 29 February 2016 /
Published online: 16 March 2016

Abstract Facial expressions are one of the most powerful, natural and immediate means for human being to communicate their emotions and intensions. Recognition of facial expression has many applications including human-computer interaction, cognitive science, human emotion analysis, personality development etc. In this paper, we propose a new method for the recognition of facial expressions from single image frame that uses combination of appearance and geometric features with support vector machines classification. In general, appearance features for the recognition of facial expressions are computed by dividing face region into regular grid (holistic representation). But, in this paper we extracted region specific appearance features by dividing the whole face region into domain specific local regions. Geometric features are also extracted from corresponding domain specific regions. In addition, important local regions are determined by using incremental search approach which results in the reduction of feature dimension and improvement in recognition accuracy. The results of facial expressions recognition using features from domain specific regions are also compared with the results obtained using holistic representation. The performance of the proposed facial expression recognition system has been validated on publicly available extended Cohn-Kanade (CK+) facial expression data sets.

✉ Joonwhoan Lee
chlee@jbnu.ac.kr

Deepak Ghimire
deepak@keti.re.kr

Sunghwan Jeong
shjeong@keti.re.kr

San Hyun Park
shpark@keti.re.kr

[1] Korea Electronics Technology Institute, Jeonju-si, Jeollabuk-do 561-844, Republic of Korea

[2] Division of Computer Engineering, Jeonbuk National University, Jeonju-si, Jeollabuk-do 561-756, Republic of Korea

Keywords Facial expressions · Local representation · Appearance features · Geometric features · Support vector machines

1 Introduction

Over the last two decade human facial expression recognition (FER) has emerged as an important research area. Facial expressions are one of the most powerful, natural and immediate means for human being to communicate their emotions and intensions. Automated and real time FER impact important applications in many areas such as human-computer interaction, health-care, driver safety, virtual reality, video-conferencing, image retrieval, human emotion analysis, cognitive science, personality development etc. Facial expressions are the one of the most important media for affect recognition. Psychologists have developed different systems to describe and quantify facial behaviors. Among them, the facial action coding system (FACS) developed by Ekman and Friesen [9] and Ekman et al. [10] is the most popular one. FACS provides a description of all possible and visually detectable facial variations in terms of 33 action units (AUs). All facial expressions can be modeled by a single AU or combination of AUs. In [3], a review of signals and methods for affective computing is presented in which most of the research for facial expression analysis are based on detection of six basic emotions defined by P. Ekman [8] namely; anger, disgust, fear, happiness, sadness and surprise. Sometimes, a neutral expression is considered as a seventh expression. In this paper we focus on recognizing the basic facial expressions.

There are two main approaches for a typical FER system: (a) Processing 2D static images, (b) Processing image sequences. In the first approach, which is more difficult than processing image sequences since less information is available, only the current frame is utilized in order to recognize the expressions (e.g. [12]). Whereas, in the second approach, the temporal information of the image sequence displaying emotion is utilized in order to recognize facial expressions (e.g. [11]). The neutral face is used as a baseline face, and FER is based on the difference between the neutral face and the succeeding input face images. Besides these 2D approaches for FER, researchers have also developed methodologies for FER from 3D mesh video. A recent survey on FER in 3D video sequences in presented in [7].

In terms of features, FER system can be categorized into two categories – appearance feature based and geometric feature based classification. Geometry-based features describe the shape of the face and its components such as mouth or eyebrow, whereas appearance-based feature describe the texture of the face caused by expression. In sequence-based method, the geometric feature primarily captures the temporal information within a sequence caused by expression such as the displacement of facial feature points between the current frame and the initial frame [11], whereas in frame-based method, the geometric features are extracted to represent shape of facial components such as the distance between fiducial points [22]. Appearance features are also utilized for the recognition of facial expression in both frame-based [12, 22] as well as in sequence-based systems [37]. The combination of appearance information and geometric information can also be utilized for FER [22].

Local Binary Pattern (LBP) and its variants are the most widely used appearance features for the recognition of facial expressions [6, 26, 36–38]. Refer to [26] for a comprehensive study on FER using LBP descriptors. Similarly, Histogram of Orientation Gradient (HOG) [12], wavelets [22, 36], Linear Discriminant Analysis (LDA) [27, 28, 38], Independent Component Analysis (ICA) [28] etc. are also widely used appearance-based feature for the recognition of facial expressions. Recently, Non-Negative Matrix Factorization (NMF) and its

variants are also widely used for the recognition of facial expressions [32, 39]. A dual subspace NMF (DSNMF) to decompose facial images into two parts: identity and facial expression parts, is proposed in [32]. R. Zhi et al. [39] proposed Graph-preserving Sparse NMF (GSNMF) algorithm for FER. GSNMF was derived from original NMF by exploiting both sparse and graph-preserving properties. In geometric feature based approach, key facial points are first detected and then tracked in case of sequence based FER. Ghimire and Lee [11] used tracking result of 52 facial key points modeled in the form of points and lines features selected using multiclass AdaBoost, and classified using SVM for the recognition of facial expressions. In [17], geometric displacement of certain selected candidate nodes, defined as the difference of the node coordinates between the first and the greatest facial expression intensity frame are used as geometric features for recognition of six basic facial expressions. A. Poursaberi et al. [22] and A. Saeed et al. [24] utilize distance between selected fiducial points as a geometric feature from single frame for FER. A novel bag-of-words based approach is recently proposed in [1] for recognizing facial expressions from a video sequence. Each video sequence is represented as a specific combination of local motion patterns captured in motion descriptors which are unique combinations of optical and image gradient. Pose invariant FER based on a set of characteristics facial points extracted using Active Appearance Models (AAMs) is presented by Rudovic et al. [23]. A Coupled Scale Gaussian Process Regression (CSGPR) model is used for head-pose normalization.

Large number of classification techniques has been employed for accurate expression recognition. Authors in [6, 11, 17, 24, 26, 37] used Support Vector Machines (SVMs), whereas, authors in [28, 33–35] utilized Hidden Markov Model (HMM) for the recognition of facial expressions. SVM is suitable for recognizing facial expressions from single frame as there is no direct probability estimation in SVMs. HMM's are mostly used to handle sequential data when frame level features are used. This has an advantage over other classifiers. Besides them, Gaussian Mixture Model (GMM) [25], and Dynamic Bayesian Networks (BN) [18] are also utilized for learning facial expressions. Recently deep learning, which integrates both feature extraction and learning procedure within deep networks, is being widely used for FER [19, 31].

In this paper, different from other approaches for FER, we use facial point locations to define a set of face regions instead of representing face as a regular grid based on face location alone, or using small patches centered at facial key point locations. By representing face in such a way we can obtain better-registered descriptors as compared to grid based representation. Local appearance features are computed based on a new definition of the face regions. The second contribution in this paper is the use of geometric features from corresponding local regions in combination with appearance features. Since, facial point locations are used to define face local regions, geometric features defines the shape of the local regions which vary according to face emotion.

The rest of the paper is organized as follows. Section 2 describes the theoretical components used in our proposed FER system. Experimental results and discussion to validate proposed FER system is presented in Section 3 and Section 4 presents concluding remarks of this paper.

2 Methods

The proposed FER system consists of four steps. First, the facial landmark positions are estimated using [15]. The detail of this technique is described in Section 2.1. Then we divide

the face region into arbitrary shaped local regions based on estimated landmark positions for better face registration. Using validation dataset of facial expressions, exhaustive search technique is employed to find important local face region, which results in feature dimensionality reduction as well as better expression recognition performance, which is the main contribution of this paper (Section 2.2). In third step (Section 2.3), appearance and geometric feature extraction is explained. Both, holistic representation and proposed local representation are used to extract appearance descriptor. The results from both representations are compared to each other. The proposed method uses features from local representation. Finally, SVM is used to learn the facial expression which is explained in Section 2.4. The overall flow diagram of the proposed system for FER is shown in Fig. 1.

2.1 Landmark position estimation

We use landmark detection method recently presented by Kazemi et al. [15] which is implemented in DLIB machine learning toolkit [16]. Accurately estimating landmark positions is very important step in the proposed system as the error in localization cumulates in the succeeding steps. This method uses ensemble of regression trees to estimate the face landmark positions directly from a sparse subset of pixel intensities, achieving super-real-time performance with high quality predictions. They present general framework based on gradient boosting for learning an ensemble of regression trees that optimizes the sum of squared error loss and naturally handles missing or partially labeled data. This method accurately estimates the landmark position not only in neutral face, but also in a face with different expressions as shown in Fig. 2.

2.2 Face local representation & region selections

In this paper, different from other approaches for FER, we use facial point locations to define a set of face regions instead of representing face as a regular grid based on face location alone or using small patches centered at facial key point locations as in [14]. Through this we can

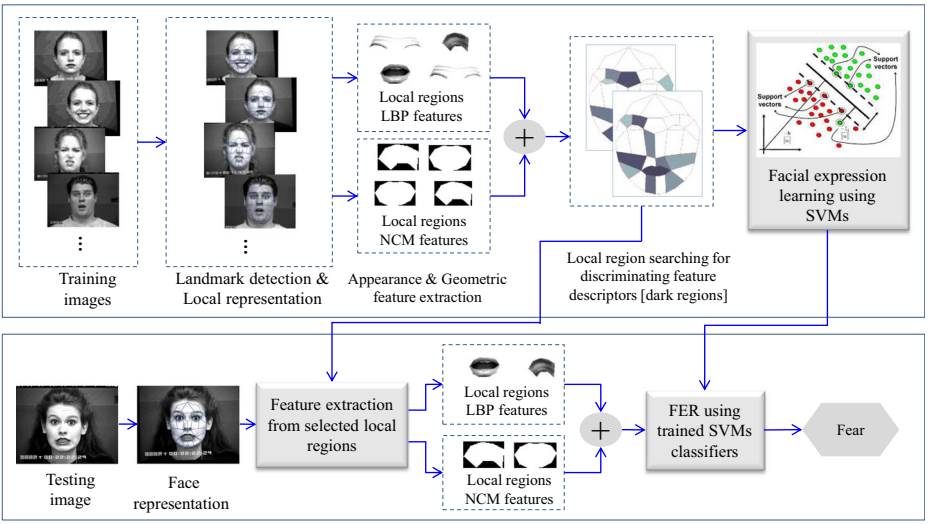

Fig. 1 Overall flow chart of the proposed facial expression recognition system

Fig. 2 Facial landmark estimation using regression tree based method presented in [15]. Note that the landmark positions in chin region are located according to eyebrow and eye landmark positions in order to cover chin region

obtain better-registered descriptors compared to grid based representation (see Fig. 3). In local method, the physical part of the face remains unchanged despite the given expression, whereas, in holistic representation the physical part of the face, from which feature descriptors are extracted, can vary depending on expression. The division of face local region is based on the expert knowledge regarding face geometry, and movement of facial muscles due to different expressions.

As shown in Fig. 4a, we divide face region into 29 local regions. Using all the 29 local regions for extracting appearance and geometric features can result in high feature dimensionality. As we know only few AUs or combination of subset of AUs contribute to producing basic facial expressions, we need not use features from all those regions for learning basic expressions. We have to select only a subset of local regions among 29 face local regions. We used exhaustive search technique using facial expression validation dataset. The starting region for searching subset of local regions is set to mouth region. The mouth region contributes most discriminating information for learning facial expressions which is validated as shown in Fig. 4b. Now only a subset of local regions is used for extracting geometric and appearance features for learning facial expressions and detection of facial expressions (Fig. 4c). Primarily, the face local region around mouth and eyes are selected as these regions carry the most discriminating information for learning facial expressions. Another interesting thing regarding those selected local regions is that only one local region from the symmetric face local regions

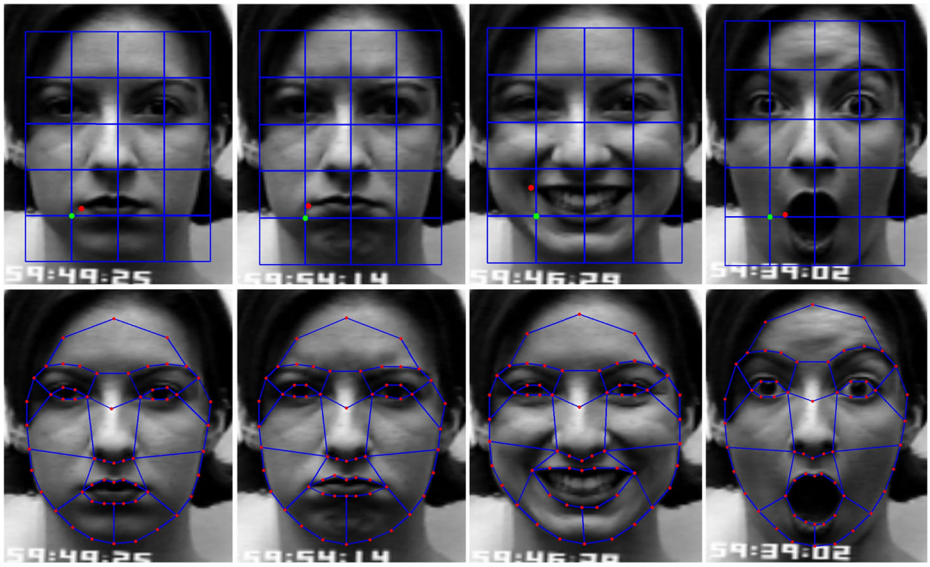

Fig. 3 Grid versus local representation (left to right; neutral, sadness, happy and surprise): regular grid (block) based representation (first row) and domain specific local region based representation (second row). Better face registration is obtained using local region based representation. Green and red dot shows the variation of registration error using block based representation in different expression

is selected which helps in removing redundant information. More details of local region selection and their effect in FER will be discussed in experimental result section.

2.3 Feature extraction

The main aim of this work is to show that the facial features descriptors for FER by dividing face region into domain specific local region outperforms feature descriptors extracted using holistic representations. We use basic LBP descriptor as appearance feature and normalized central moments (NCM) descriptors as geometric feature descriptors, each of them are

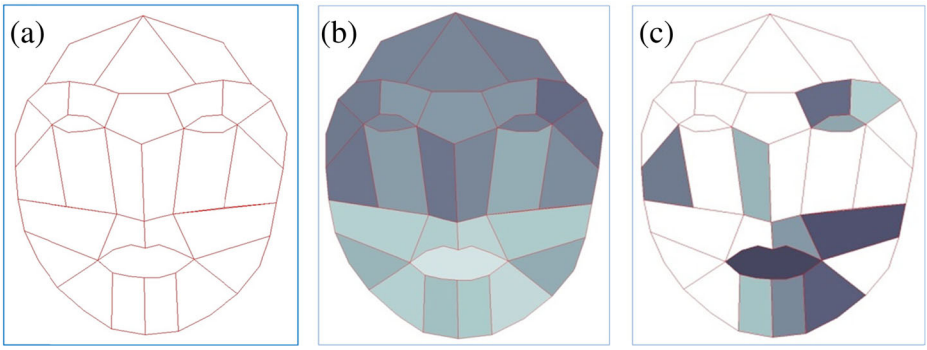

Fig. 4 Local region specific face representation and region selection. **a** Division of face image into 29 regions based on expert knowledge regarding AUs and facial expressions. **b** Labeling of face local regions according to expression recognition using feature descriptors form individual face local regions (for ex., mouth region (dark–light) contributes the most information for discriminating facial expression). **c** Selected local regions (using validation set) form which feature descriptors will be used for recognizing facial expressions

explained briefly in the following subsections. Finally, we concatenate LBP and NMC features to feed into SVM machine learning algorithm.

2.3.1 Local binary pattern

Different appearance features such as LBP [6, 26, 36–38], HOG [12], Local Gabor Binary Pattern (LGBP) [30], Scale Invariant Feature Transform (SIFT) [29] etc. are widely used by many researchers for FER. Since the local regions are of varying size and shapes we can only use histogram-based feature descriptors; therefore in our system we choose LBP feature as appearance feature. The whole face region is divided into region specific local regions as shown in Fig. 4a. The feature descriptors for FER are used only from subset of local regions detected using exhaustive search technique.

In LBP operator [20], a binary code is produced for each pixel in an image by thresholding its neighborhood with the value of the center pixel. It was originally defined for 3×3 neighborhoods giving 8 bit codes based on the 8 pixels around the center pixel. The operator was later extended to use neighborhood of different sizes, image planes, rotation invariant LBP etc. In our system we just use the basic LBP operator. The operator labels the pixels of an image by thresholding a 3×3 neighborhood of each pixel with the center value and considering the result as a binary number. A 256-bin histogram of the LBP labels is computed over a region and is used as a texture descriptor. A LBP is 'uniform' if it contains at most one 0–1 and one 1–0 transition when viewed as a circular bit string. For instance, 00000000, 00111100 and 11000001 are uniform pattern. It is observed that uniform patterns account for nearly 90 % of all patterns in the (8, 1) neighborhood in texture images.

After labeling an image with LBP operator, a histogram of the labeled image $f_l(x, y)$ can be defined as:

$$H_i = \sum_{x,y} I(f_l(x,y) = i), i = 0, 1, \ldots, n-1 \tag{1}$$

where n is the number of different labels produced by the LBP operator and,

$$I(A) = \begin{cases} 1, & \text{A is true} \\ 0, & \text{A is false} \end{cases} \tag{2}$$

We use uniform LBP pattern in our system from each selected local region. Figure 5 gives an example of the basic LBP operator.

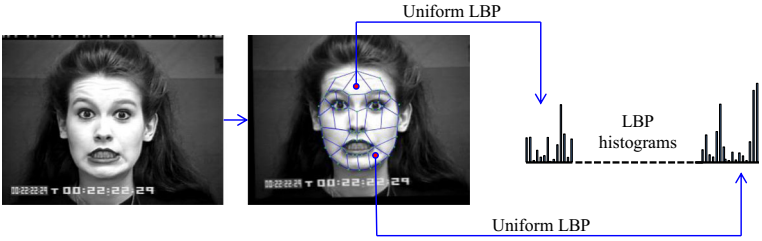

Fig. 5 LBP feature extraction process

The LBP features are also extracted using grid representation. The total face region is divided into regular grids (Fig. 3 first row) and the result of FER from grid representation is compared with the result from proposed local representation.

2.3.2 Normalized central moments

Movement of facial landmarks or special positions of facial landmarks are used by many researchers to extract geometric information for this particular problem [11, 17, 22, 24]. In our system, movement of facial landmarks cannot be used as it is a frame based system. The shape and size of local regions in our representation varies for different expressions, therefore we also want to capture shape information as geometric feature descriptor. The normalized central moments up to three orders are used from each selected local regions in our face representation which is calculated as follows.

The spatial moments (m_{ji}) are computed as,

$$m_{ji} = \sum_{x,y} \left(I(x,y).x^j.y^i \right) \tag{3}$$

where $I(x,y)$ is the binary image with face local shape represented with 1 and background with 0.

The central moments (mu_{ji}) are computed as,

$$mu_{ji} = \sum_{x,y} \left(I(x,y).\left(x-\overline{x}\right)^j.\left(y-\overline{y}\right)^i \right) \tag{4}$$

where $(\overline{x},\overline{y})$ is the mass center.

$$\overline{x} = \frac{m_{10}}{m_{00}}, \quad \overline{y} = \frac{m_{01}}{m_{00}} \tag{5}$$

The normalized central moments (nu_{ji}) are now computed as:

$$nu_{ji} = \frac{mu_{ji}}{m_{00}^{(i+j)/2+1}} \tag{6}$$

Thus obtained geometric descriptors are concatenated with appearance descriptors and FER is performed using SVM classification.

2.4 Support vector machines for FER

SVMs are powerful tool for both binary and multi-class classification and regression. SVMs are robust against outliers. For two-class classification SVM estimates the optimal separating hyper-plane between the two classes by maximizing the margin between the hyper-plane and closest points of the classes. In case of binary classification task with training data $x_i (i=1,\ldots,N)$ and corresponding class labels $y_i = \pm 1$, the decision function can be formulated as

$$f(x) = \text{sign}\left(w^T x + b\right) \tag{7}$$

where $w^T x + b = 0$ denotes a separating hyper-plane, b is the bias or offset of the hyper-plane from the origin in the input space, and w is a weight vector normal to the separating hyper-plane. The region between hyper-planes is called margin band,

$$\gamma = \frac{2}{\|w\|} \tag{8}$$

where $\|w\|$ denotes 2-norm of w. Finally, choosing the optimal values (w, b) is formulated as a constrained optimization problem, where (8) is maximized subject to the following constrain:

$$y_i\left(w^T x_i + b\right) \geq 1 \quad \forall i \tag{9}$$

Several one-versus-all SVM classifiers are used to handle the multiclass expression recognition problem. In our system we use a publicly available implementation of SVM, *libsvm* [4], with radial basic function (RBF) kernel. The optimal parameter selection is performed based on the grid search strategy [13]. OpenCV [2] implementation of *libsvm* is used in our experiment.

3 Experimental results

The performance of the proposed FER system is evaluated in publicly available extended Cohn-Kanade (CK+) facial expression dataset [21]. CK+ dataset consists of sequence of images to represent single expression, which starts from neutral face and evolves to peak facial expression intensity. We use only the peak expression frames from each sequence to validate the performance of our proposed FER system. A five-fold cross validation was used to make maximum use of the available data. The classification accuracy is the average accuracy across all five trails. The confusion matrices are given to get better picture of the FER accuracy.

CK+ dataset consists of 593 sequences from 123 subjects. Each image sequence starts with onset (neutral expression) and ends with a peak expression (last frame). The peak expression is fully coded by FACS. Only 327 of the 593 sequences were given label for the human facial expressions; this is due to these are the only ones that fit the prototypic definition. We used two peak expression frames for anger, fear and sadness expression, where as we used single peak expression frame for disgust, happy and surprise expression. This is due to anger, fear and sadness expression have little amount of sequences as compared with the rest of the expressions. We perform experiment for six-class FER as well as seven-class FER which also includes neutral expression. Total of 60 neutral frames are chosen; which is the first frame in the expression sequence. As a result we have 60 frames for neutral (NU), 88 frame for anger (AN), 62 frames for disgust (DI), 54 frames for fear (FE), 69 frames for happiness (HA), 64 frames for sadness (SA), and 81 frames for surprise (SU) expressions.

3.1 Grid versus proposed local representation based FER

At first, we use features from holistic representation (regular grid) for FER. Uniform LBP descriptors are extracted by dividing face image into 4×5 blocks and 5×6 blocks, and LBP features form each block are concatenated and used for FER. Figure 6a, b show the confusion matrices for FER using grid representation with 4×5 blocks, and 5×6 blocks, respectively.

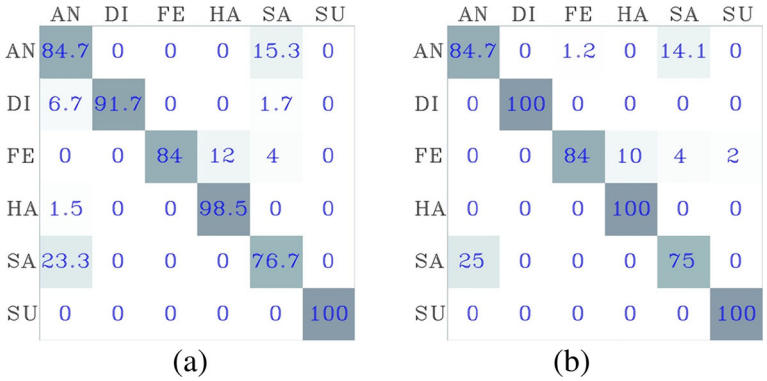

Fig. 6 Confusion matrices of FER using LBP feature descriptors from grid representation: **a** 4 × 5 grid; **b** 5 × 6 grid

The average recognition accuracy using 5 × 6 blocks is 90.62 % which is slightly higher than using 4 × 5 blocks, i.e., 89.25 %. The main aim of this paper is to show that proposed local representation outperforms grid based representation rather than competing with the accuracies in the literature, therefore we only experimented with LBP features for both representation and did not explore the performance of other appearance features.

In the proposed local region based representation, we first divide the whole face region into 17 local regions (see Fig. 3, second row). The single uniform LBP is extracted from each local regions in addition to whole face region covered by outer boundary resulting in 18 LBP histograms. They are now concatenated to get 1062 (59 × 18) dimensional feature vector. Further, we divide the whole face region into 29 local regions (see Fig. 4) and calculate LBP histogram from each region as well as from whole face region, resulting in 1770 (59 × 30) dimensional feature vector. Figure 7a and b show the confusion matrices using LBP features extracted by dividing face into 17 and 29 local regions. An improvement of ~2 % in FER accuracy is achieved with dense local region representation, i.e., average of 91.37 % accuracy using 17 local regions, and average of 93.60 % accuracy using 29 local regions. So, the rest of the experiments in this paper are performed using face representation with 29 local regions.

Our second contribution in this paper is the use of geometric shape features from the proposed local representation. The NCM up to 3rd order using Eq. 3 to Eq. 6 are calculated from each of the 29 local regions plus the whole face region, resulting in 210 (30 × 7) dimensional feature vectors, which is quite low in dimension as compared to LBP descriptor. Figure 7c shows the confusion matrix for FER using NCM features from proposed local face representation. An average of 89.67 % of recognition accuracy is achieved using NCM features.

Finally, we concatenate appearance feature and geometric features into single feature vector. Figure 7d shows the confusion matrix for FER using proposed feature set and local face representation. An average recognition accuracy of 94.83 % is achieved using combination of LBP and NCM features from proposed local face representation. Individual recognition accuracy of expressions anger, fear and sadness is in the range of 90 % whereas disgust, happiness and surprise are recognized with very high accuracy. Figure 8 shows the comparison

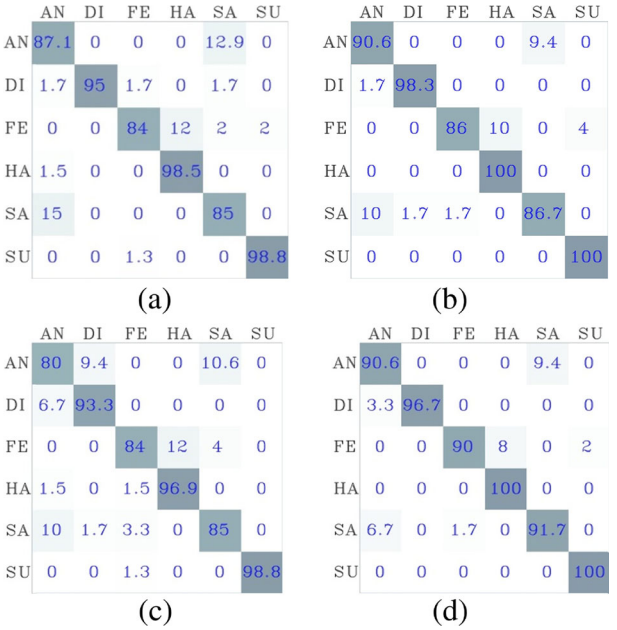

Fig. 7 Confusion matrices for FER using LBP feature descriptors [**a** 17 local regions (see Fig. 3, second row); **b** 29 local regions (see Fig. 4)], NCM shape features [**c** 29 local regions (see Fig. 4)] and LBP + NCM features [**d** 29 local regions (see Fig. 4)] from proposed local region representations

of FER accuracies using grid versus proposed local representation. In our experiment we extract single LBP histogram from each local region. Therefore the LBP feature dimensionality using 5×6 grid is equal to the LBP feature dimensionality from 29 local regions plus whole face region given by outer boundary. But the classification accuracy is better with local representation. This is mainly due to better face registration in local representation as compared to grid representation. This proves that the proposed face representation can produce better classification accuracy of facial expressions as compared to traditional grid based representation.

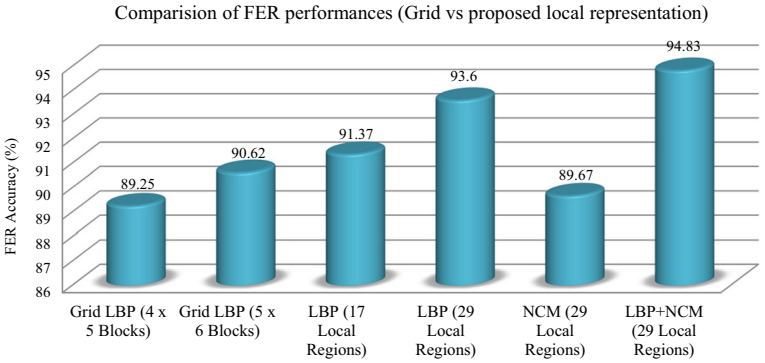

Fig. 8 Comparison result of FER accuracies using grid versus proposed local region based representations

3.2 Selection of important local regions for FER

In the proposed FER system, whole face region is divided into 29 local regions. The division of face region is based on landmark locations, therefore each local region are well registered regardless of facial expression. In contrast, in grid representation there will be registration error due to different expressions. As we have better registration of each local region, we do not require all 29 local regions in order to discriminate basic facial expressions. As face has symmetric property, most of the region provides redundant information. Therefore we used exhaustive search scheme to search for important local regions as search space is not big enough to use complex search algorithms. We use mouth region as seed for searching other regions because mouth region provides most discriminating information for recognizing facial expressions (see Fig. 4b). The validation set from CK+ dataset is used to search for other local regions. Figure 9 shows the result of local region selection. In each step, a new local region which contributes in getting highest recognition accuracy is added. Total of 13 local regions are selected out of the 29 local regions using LBP as a feature descriptor.

As we can see regions are selected from mouth, chin, eyes, eye-brow areas, which contribute the most on discriminating basic facial expressions. For instance, as we know there is large movement of muscles around mouth region in happy and surprise expressions and small movements in the case of other expressions. In case of disgust and fear, there is muscle contraction around eye region especially in between eye brows. Most of the face symmetric local regions are not selected as they carry redundant information. Therefore the selected local regions will carry sufficient information for learning basic facial expressions.

Figure 10 shows the confusion matrices for FER using feature descriptors extracted only from the selected local regions. One interesting point we observe from this experiment is that, the accuracy increases even though we use only the selected local regions instead of all local regions

Fig. 9 Search of local regions for getting highest accuracy of FER using LBP descriptors. Starting form mouth region in each step a new region is added using exhaustive search scheme which contributes the most. Dark region contributes more information and the light dark region contributes less information for FER

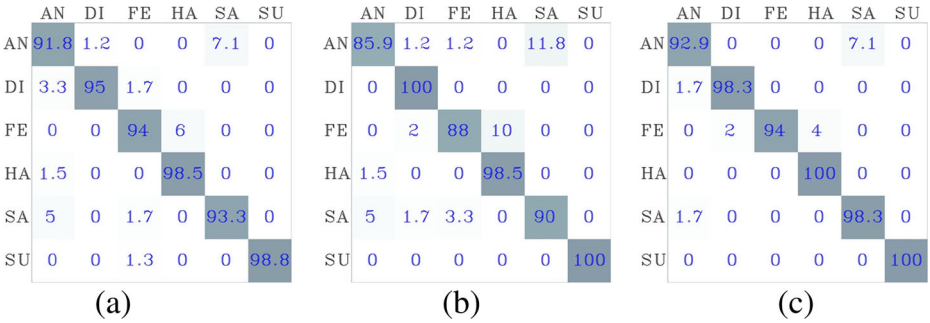

Fig. 10 Confusion matrices for FER using feature from selected local regions; **a** LBP, **b** NCM, and **c** LBP + NCM

from whole face area. On the other hand the dimensionality of feature descriptor also decreases approximately by half. Using only the NCM shape features we obtained 93.72 % recognition accuracy, using LBP feature descriptor we obtained 95.22 % recognition accuracy, whereas after concatenating both the features we obtained 97.25 % recognition accuracy. This increment in accuracy shows that not all the face local regions provide discriminating information for learning facial expressions. Therefore region selection works as filtering of face local region which affects the learning of facial expressions.

Another major advantage of the proposed local region selection based FER is the reduction in computational complexity of the algorithm. The region selection is an offline process. In our experimental setup among 29 local regions we selected a subset of 13 local regions. As the search space is not big, we used exhaustive search scheme for region selection. Now, once we have selected local regions, during FER there are mainly three stages: facial key point detection, feature extraction from selected local regions and finally facial emotion classification as shown in Fig. 1. As we use super-fast implementation of ensemble of regression tree based face alignment method proposed in [7], the whole FER process runs in real time.

3.3 FER including neutral expression

Sometimes neutral expression is considered as seventh expression. Methods for discriminating neutral frame before emotion classification have been proposed (e.g. [5]). If we can distinguish neutral frame in the early stage, processing each and every frame to classify emotions is not necessary, as user stays neutral most of the time [5]. However, on the other hand we can consider neutral frame as seventh expression and perform seven-class emotion classification. Discarding neutral frame in the early stages requires algorithm with high accuracy otherwise the error will accumulate in the next stage.

In our study we also performed experiments considering neutral face as seventh expression. Figure 11 shows the confusion matrices for FER using feature descriptors only from selected local regions after including neutral expression. Figure 12 shows the graph of FER performance comparison using different feature extraction techniques for both six-class and seven-class expressions. Performance of the seven-class expression recognition decreases as compared to six-class expression recognition. Although, the performance of the proposed local region based FER is better as compared to grid representation based FER, the performance decreases while including neutral expression. This is mainly due to anger and sadness expressions are highly confused with neutral expressions which can be seen in Fig. 11. In our experiment anger and sadness are classified with

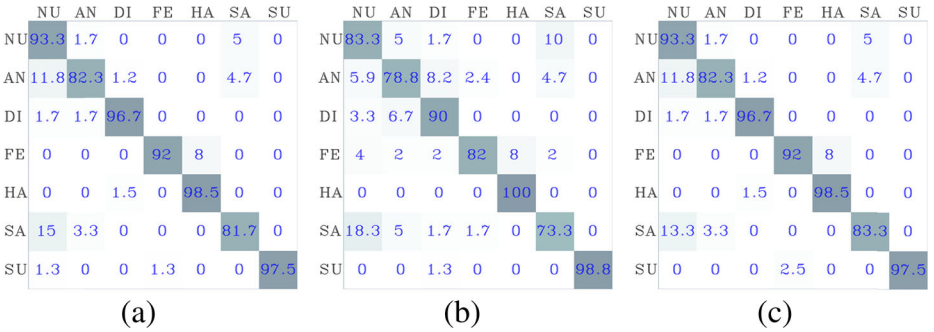

Fig. 11 Confusion matrices for FER using feature from selected local regions including neutral expression; **a** LBP, **b** NCM, and **c** LBP + NCM

92.9 % and 98.3 % accuracy, respectively, after dropping neutral expression whereas the performance decreases to 82.3 % and 83.3 % for anger and sadness, respectively, after including neutral expression. This makes sense because if we see sample expression frames in Fig. 3, there is very little difference between above specified facial expressions. Therefore it is better to drop neutral expression in the early stage as in [5] before performing emotion classification.

3.4 Comparison with state-of-the-art FER techniques

The results of our proposed method are compared with several results of FER proposed by different researchers in the literature. In our experiment we achieved 91.95 % and 97.25 % recognition accuracy for seven and six classes, respectively, FER using combination of basic LBP feature and NCM shape feature from selected facial local regions of single frame. In [12], authors achieved 97.3 % of recognition accuracy for seven classes of facial expressions using ensemble of 15 extreme learning machine (ELM) classifiers. Although this result seems better as compared to ours, the classification system is complex and they did not perform the cross validation test. A. Poursaberi et al. [22] achieved 92.02 % of recognition accuracy using texture and geometric features from single facial image. In [36], 95.24 % of recognition accuracy is obtained for seven class expressions using SVM classifier on 2478 dimensional LBP features from 6 × 7 facial grids. Although the facial

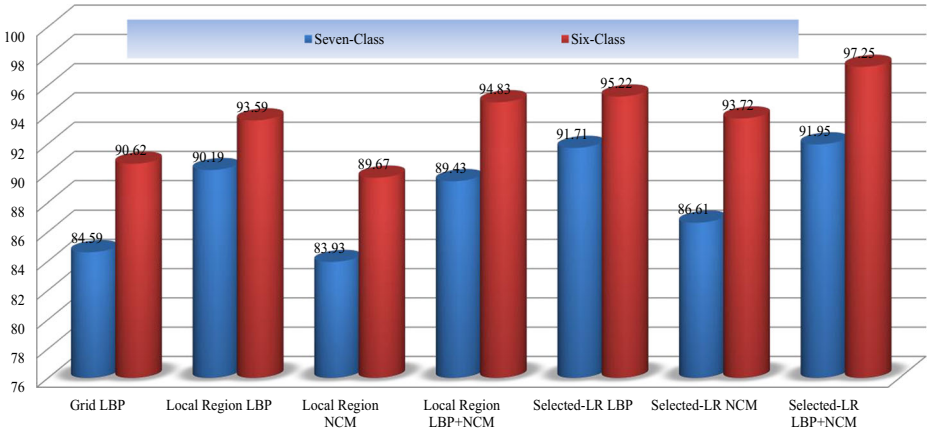

Fig. 12 FER performance comparison using different feature extraction techniques

Table 1 FER performance comparison in CK+ database with different methods in the literature

Reference	Method	Image data	Class	Accuracy (%)
[12]	HOG feature, ELM Ensemble	Frame	7	97.30
[11]	AdaBoost selected geometric features	Sequence	6	97.35
[22]	Texture and geometric features	Frame	6	92.20
[37]	LBP-TOP + VLBP	Sequence	6	96.26
[38]	LBP + KDIsomap	Frame	7	94.88
[36]	LBP + SVM	Frame	7	95.24
[27]	Stepwise LDA + hidden conditional random field	Frame	6	96.83
[39]	Graph-preserving sparse NMF	Frame	6	94.30
[17]	Geometric key displacement features	Sequence	6	99.70
[24]	Geometric features	Frame	7	83.01
[34]	Enhanced independent component + FLDA	Frame	6	93.23
[18]	Geometric features + dynamic Bayesian network	Sequence	6	94.04
Ours	Local representation, LBP + NCM features	Frame	6	97.25
		Frame	7	91.95

features and classifier is same as in our case the experimental setup seems different as the number of frames for each expression is different. For instance, they used 32 anger, 116 neutral images as opposed to 88 anger, and 60 neutral images in our experiment. The sequence based FER system in [17] achieved 99.7 % of recognition rate using key point displacement features from neutral to peak expression, which is the highest accuracy so far. In [24], 83.01 % of recognition rate is achieved using distance based features extracted from 8 facial key points from single face image. Another sequence based method proposed by Ghimire and Lee [11], achieved 97.35 % recognition accuracy using geometric feature descriptors. Literature in expression recognition shows that sequence based technique achieved slightly higher recognition accuracy as compared to single frame based techniques. But the major difficulty with sequence based technique is that identification of the image frame at which expression actually starts evolving, as well as identification of the peak expression frame, or the identification of time duration at which facial expression actually occurred. Table 1 shows the comparison of FER performance with different methods in the literature. The proposed method achieved competitive recognition accuracy even using single facial frame as compared to sequence based methods in the literature.

4 Conclusions

In this paper we propose a new approach for FER that uses a combination of appearance and geometric shape features from local face regions. The proposed face representation provides better face registration than mainstream face representation, i.e. holistic representation. Performance improvement as well as dimensionality reduction is obtained with searching local face regions carrying the most discriminating information for facial expression classification. Several experiments were performed in the CK+ dataset in order to validate the usefulness of the proposed FER technique for both six-class and seven-class expressions. We compared the proposed local region based representation technique with grid based holistic representation. Experimental results showed that the local region based representation performs better as compared to holistic representation.

Even though the main aim of the paper was not to focus on competing in terms of accuracy in the literature, we obtained comparative and at times better result of FER using proposed technique as compared to the different techniques in the literature. We believe that, there is a room for performance improvement by searching best features for discriminating facial expressions within proposed framework. Our future work will focus in the same direction.

References

1. Agrawal S, Mukherjee DP (2015) Facial expression recognition through adaptive learning of local motion descriptors. Multimed Tools Appl. doi:10.1007/s11042-015-3103-6, online first
2. Bradski G (2000) "The OpenCV library," Dr. Gobb's. J Softw Tools
3. Calvo RA, D'Mello S (2010) Affect detection: an interdisciplinary review of models, methods and their applications. IEEE Trans Affect Comput 1(1):18–37
4. Chang C-C, Lin C-J (2011) LIBSVM: a library for support vector machines. ACM Trans Intell Syst Technol, pp. 2:27:1—27:27. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm
5. Chiranjeevi P, Gopalakrishna V, Moogi P (2015) Neutral face classification using personalized appearance models for fast and robust emotion detection. IEEE Trans Image Process 24(9):2701–2711
6. Cruz AC, Bhanu B, Thakoor NS (2014) Vision and attention theory based sampling for continuous facial emotion recognition. IEEE Trans Affect Comput 5(4):418–431
7. Danelakis A, Theoharis T, Pratikakis I (2014) A survey on facia expression recognition in 3D video sequences. Multimed Tools Appl 74(15):5577–5615
8. Ekman P (1989) "The argument and evidence about universals in facial expressions of emotions," Handbook of Social Psychophysiology. Wiley, Chichester, pp 143–164
9. Ekman P, Friesen W (1978) Facial action coding system (FACS). Consult. Psychol. Press, Palo Alto
10. Ekman P, Friesen WV, Hager JC (2002) Facial action coding system. A Human Face, Salt Lake City
11. Ghimire D, Lee J (2013) Geometric feature-based facial expression recognition in image sequences using multi-class AdaBoost and support vector machines. Sensors 13:7714–7734
12. Ghimire D, Lee J (2014) Extreme learning machine ensemble using bagging for facial expression recognition. J Inf Process Syst 10(3):443–458
13. Hsu CW, Chang CC, Lin CJ (2010) A practical guide to support vector classification. Technical Report; Department of Computer Science, National Taiwan University, Taiwan
14. Jiang B, Martinez B, Valster MF, Pantic M (2014) Decision level fusion of domain specific regions for facial action recognition. Int. Conf. on Pattern Recog., p 1776–1781, 24–28
15. Kazemi V, Sullivan J (2014) One millisecond face alignment with an ensemble of regression trees. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Columbus, 23–28 June 2014, p 1867–1874
16. King DE (2009) Dlib-ml: a machine learning toolkit. J Mach Learn Res 10:1755–1758
17. Kotisa I, Pitas I (2007) Facial expression recognition in image sequence using geometric deformation features and support vector machines. IEEE Trans Image Process 16(1):172–187
18. Li Y, Wang S, Zhao Y, Ji Q (2013) Simultaneous facial feature tracking and facial expression recognition. IEEE Trans Image Process 22:2559–2573
19. Liu P, Han S, Meng Z, Tong Y (2014) Facial expression recogniton via a boosted deep belief network. In: Proc. IEEE Conf. on CVPR, p 1805–1812, 23–28
20. Ojala T, Pietikainen M, Maenpaa T (2002) Multiresolution gray scale and rotation invarient texture analysis with local binary patterns. IEEE Trans Pattern Anal Mach Intell 24:971–987
21. Pantic M, Valster M, Rademaker R, Maat L (2010) The extended Cohn-Kanade dataset (CK+): a complete dataset for action unit and emotion –specific expressions. In: Proc. of 3rd IEEE Workshop on CVPR for Human Communicatin Behaviour Analysis, p 94–101
22. Poursaberi A, Noubari HA, Gavrilova M, Yanushkevich SN (2012) Gauss-Laguerre wavelet textural feature fusion with geometrical information for facial expression identification. EURASIP J Image Video Process 2012:17, pp. 1–13
23. Rudovic O, Pantic M, Patras I (2013) Coupled Gaussian processes for pose-invarient facial expression recognition. IEEE Trans Pattern Anal Mach Intell 35(6):1357–1369
24. Saeed A, Al-Hamadi A, Niese R, Elzobi M (2014) Frame-based facial expression recognition using geometric features. Adv Hum Comput Interact 2014:1–13

25. Schels M, Schwenker F (2010) A multiple classifier system approach for facial expressions in image sequence utilizing GMM Supervectors. In: Proc. of the 2010 10th Int. Conf. on Pattern Recog., p 4251–4254
26. Shan C, Gong S, McOwan PW (2009) Facial expression recognition based on local binary patterns: a comprehensive study. Image Vis Comput 27:803–816
27. Siddiqi MH, Ali R, Khan AM, Park Y-T, Lee S (2015) Human facial expression recognition using stepwise linear discriminant analysis and hidden conditional random fields. IEEE Trans Image Process 24(4):1386–1398
28. Siddiqi MH, Lee S, Lee Y-K, Khan AM, Truc PTH (2013) Hierarchical recognitoin scheme for human facial expression recognition systems. Sensors 13:16682–16713
29. Soyel H, Demirel H (2011) Improved SIFT matching for pose robust facial expression recognition. In: Prof. IEEE Int. Conf. on FG, p 585–590, 21–25
30. Sun X, Xu H, Zhao C, Yang J (2008) Facial expression recognition based on histogram sequence of local Gabor binary patterns. In: Proc. IEEE Conf. on Cybernatics and Intell. Systems, p 158–163, 21–24
31. Susskind JM, Hinton GE, Movellan JR, Anderson AK (2008) Generating facial expressions with deep belief nets. In: Kordic V, (ed) Affective computing, emotion modeling, synthesis and recognition, p 421–440
32. Tu Y-H, Hsu C-T (2012) "Dual subspace nonnegative matrix factorization for person-invarient facial expression recognition," 21st Int. Conf. on Pattern Recognition (ICPR 2012), p 2391–2394
33. Uddin MZ, Hassan MM (2013) A depth video-based facial expression recognition system using radon transforn, generalized discriminant analysis, and hidden Markov model. Multimed Tools Appl 74(11):3675–3690
34. Uddin MZ, Lee J, Kim T (2009) An enhanced independent component-based human facial expression recognition from videos. IEEE Trans Consum Electron 55(4):2216–2224
35. Yeasin M, Bullot B, Sharma R (2006) Recognition of facial expressions and measurements of levels of interest from videos. IEEE Trans Multimedia 8(3):500–508
36. Zhang S, Zhao X, Lei B (2012) Robust facial expression recognition via compressive sensing. Sensors 12:3747–3761
37. Zhao G, Pietikäinen M (2007) Dynamic texture recognition using local binary patterns with an application to facial expressions. IEEE Trans Pattern Anal Mach Intell 29(6):915–928
38. Zhao X, Zhang S (2011) Facial expression recognition based on local binary pattern and kernel discriminant isomap. Sensors 11:9573–9588
39. Zhi R, Flierl M, Ruan Q, Kleijn WB (2011) Graph-preserving sparse nonnegative matrix factorization with applications to facial expression recognition. IEEE Trans Syst Man Cybern B Cybern 41(1):38–52

Deepak Ghimire received undergraduate degrees in Computer Engineering from Pokhara University, Nepal in 2007. He received M.S. degree and Ph.D. degree in Computer Science and Engineering from Chonbuk National University, South Korea in 2011 and 2014 respectively. He is currently a Researcher in IT Application Research Center of KETI (Korea Electronics Technology Institute), South Korea. His main research interests include image processing, computer vision, machine learning, and biometric information processing.

SungHwan Jeong received undergraduate degrees in Computer Engineering from Jeonju University, South Korea in 2004. He received M.S. degree in Biomedical Engineering and Ph.D. degree in Computer Science and Engineering from Chonbuk National University, South Korea in 2006 and 2012 respectively. He is currently a Researcher in IT Application Research Center of KETI (Korea Electronics Technology Institute), South Korea. His current research interests include image processing, computer vision, embedded vision engineering.

Joonwhoan Lee received undergraduate degrees in Electronic Engineering from the Hanyan University in 1980. He received M.S. degree in Electrical and Electronic Engineering from KAIST (Korea Advanced Institute of Science and Technology) University in 1982 and Ph.D. degree in Electrical and Computer Engineering from University of Missouri, USA in 1990. He is currently a Professor in the Department of Computer Engineering at Chonbuk National University, South Korea. His research interests include image processing, computer vision, emotion engineering.

Sang Hyun Park received his undergraduate education in Computer Science Engineering from Hankuk University of Foreign Studies, South Korea in 2000. He received M.S. degree in Computer Science Engineering from Hankuk University of Foreign Studies, Korea in 2002. Now he is Managerial Researcher in Jeonbuk Embedded System Research Center of KETI (Korea Electronics Technology Institute), Korea. His current research interests include automotive network, embedded system engineering.