Co-Clustering to Reveal Salient Facial Features for Expression Recognition

Sheheryar Khan, *Student Member, IEEE*, Lijiang Chen, *Member, IEEE*, and Hong Yan, *Fellow, IEEE*

Abstract—Facial expressions are a strong visual intimation of gestural behaviors. The intelligent ability to learn these non-verbal cues of the humans is the key characteristic to develop efficient human computer interaction systems. Extracting an effective representation from facial expression images is a crucial step that impacts the recognition accuracy. In this paper, we propose a novel feature selection strategy using singular value decomposition (SVD) based co-clustering to search for the most salient regions in terms of facial features that possess a high discriminating ability among all expressions. To the best of our knowledge, this is the first known attempt to explicitly perform co-clustering in the facial expression recognition domain. In our method, Gabor filters are used to extract local features from an image and then discriminant features are selected based on the class membership in co-clusters. Experiments demonstrate that co-clustering localizes the salient regions of the face image. Not only does the procedure reduce the dimensionality but also improves the recognition accuracy. Experiments on CK plus, JAFFE and MMI databases validate the existence and effectiveness of these learned facial features.

Index Terms—Co-clustering, facial expression recognition, feature selection, gabor wavelets, support vector machines (SVMs)

1 Introduction

Facial expressions are strong manifestation of the emotion state of a person and offer a vital behavioral measure for the learning of intention, cognitive activity and social interaction [1]. Recognizing facial expression automatically can help apprehend the psychopathology of a person in non-intrusive manner. Motivated by this significant characteristic of instantly conveying nonverbal communication, facial expression recognition plays an intrinsic role in developing the human computer interaction (HCI) and social computing fields. With the advent and availability of low cost computing and imaging devices, automatic facial expression recognition system (FER) has attracted attentions in several day-to-day application areas such as interactive video, explicit customer feedback, mimetic robots, and fatigue detection.

A study on psychophysical behavior shows that universal facial expressions across all cultures are reflected by the same emotions [2]. Based on this, present FER systems endeavor to recognize the most common set of prototype emotions namely happiness, fear, sadness, anger, surprise and disgust [50], [51], [53]. Humans are capable of recognizing these emotional states of other people naturally and effortlessly irrespective of age, gender and ethnicity. However, it is still a challenging task to learn emotions automatically with a computer at least partly due to failure to extract prominent features. These expressions are invoked by the stimulation of face muscles that are located around the eyes, nose and mouth. The facial activity corresponding to each expression can be described by certain action units (AUs) [19]. For example, the expression of surprise can be decomposed by the stretching mouth and/or raising the eyebrow as shown in Fig. 1. The occurrence of AUs around these facial parts truly indicates the saliency of these regions. However, learning this information to address the facial expression recognition task has been seldom addressed in the computer vision community.

Most existing works on appearance-based methods focus on Gabor-wavelet representations due to their promising performance and robustness against in-plane rotations and misalignments [5], [6], [7], [8]. In appearance-based methods, usually a whole face region or a pre-defined facial part is considered to extract the texture information. As reported in [49], [51], [17], dividing the face region into numerous non-overlapping blocks and then extracting features based on the statistical significance of each block enhance the recognition performance. However, accurate face alignment and the choice of the size, number and location of blocks have a direct influence on the performance of the FER system.

The process of feature selection may be performed in FER to have a semantic interpretation, such as revealing the salient spatial regions of interest and reduce the data dimensionality. In this paper, we introduce a novel algorithm to search for the most distinguished facial regions. Most of the traditional feature selection approaches based on feature ranking assess the significance of each feature individually and select features one by one [39], [40], [41]. A limitation of these methods is that the block structures in the input data and the spatial correlation among the features are neglected.

In our work, we explore the block structures in the facial expression feature matrix and propose a co-clustering based feature selection strategy. We first compute the Gabor wavelet

- *The authors are with the Department of Electronic Engineering, City University of Hong Kong, Kowloon, Hong Kong.*
  E-mail: shehekhan2-c@my.cityu.edu.hk, {lijichen, h.yan}@cityu.edu.hk.

Manuscript received 30 Dec. 2016; revised 2 Nov. 2017; accepted 2 Dec. 2017. Date of publication 11 Dec. 2017; date of current version 29 May 2020.
(Corresponding author: Sheheryar Khan.)
Recommended for acceptance by H. Gunes.
Digital Object Identifier no. 10.1109/TAFFC.2017.2780838

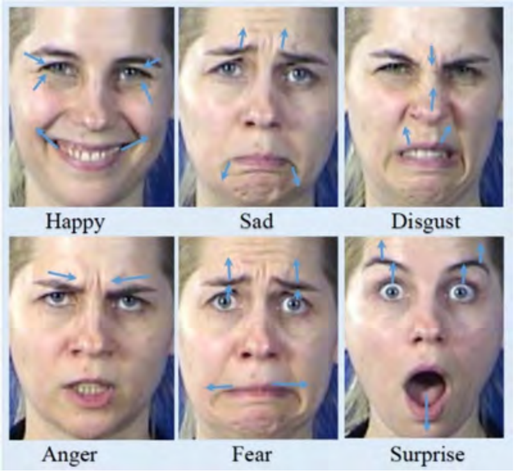

Fig. 1. An illustration of Action Units (AUs) on face images for six prototype expressions. The arrows on the images represent facial muscle movements at the AUs.

image representation of training images. We consider the problem of co-clustering of facial expression samples and the Gabor wavelet features using singular value decomposition (SVD). The decomposition offers a low-rank approximation of high dimensional data, where co-clusters are much easier to identify even in the presence of noise. Local structures in the matrix are uncovered by performing clustering with the left and right singular vectors corresponding to the largest singular values. Motivated by the fact that samples from the same block tend to share the same sparsity pattern in the low dimensional representation, we then performed feature subset selection to exploit the group information.

The novelty of the proposed method is that it selects subsets of features from the learned submatrices representing the co-occurence of samples and features. Our method takes into account all features obtained using co-clustering rather than relying on one-way clustering of the features, ranking them individually or combining the original features through a transformation. Essentially, it searches for a partition representing the bidirectional local configuration. Thus, our approach is able to group subsets of samples and subsets of features lying on a manifold.

Following the sparsity in the data matrix, the proposed methodology is able to find the most distinguished features based on the class probabilities in the learned co-clusters. In our method, only those co-clusters that have enough samples from all classes are retained for feature subset selection and remaining co-clusters are treated as noise. The feature subsets acquired during this training process localize the face regions around the eyes, mouth and nose, which are more salient in terms of distinguishing the facial expressions, as shown in Fig. 1.

We have evaluated our co-clustering algorithm on facial expression image databases. After the selection of Gabor wavelet features, we found that the selected features are accurate in terms of expression recognition rate. By varying the numbers of co-clusters, the size of selected features can be changed. The reduced feature sets are used to classify the seven basic expressions, using the multiclass support vector machine (SVM) and K-nearest neighbor (K-NN) classifiers. Our experiments revealed that the co-clustering approach offers the discriminant and low dimensional feature space for Gabor wavelet image representation. The selected features not only lead to a better recognition rate but also reduce the computational time complexity significantly.

The rest of the paper is organized as follows. Section 2 presents the review of related work. Section 3 discusses the FER framework of our proposed system, including feature extraction, and co-clustering. Experiment results are discussed in Section 4. Finally, Section 5 presents the conclusion.

2 Related Work

Human facial expression recognition has long been studied in computer vision [11]. Current systems show an encouraging continued progress, and a thorough survey of existing approaches can be found in [12]. In this section, we briefly review previous work related to facial feature extraction, expression recognition, and co-clustering.

2.1 Facial Representation

In general, the feature extraction procedure in FER systems can be categorized into geometric feature-based methods and appearance based ones [3]. Geometric feature based methods target the shapes and physical locations of facial landmarks, which are then extracted to model the face geometry [20], [23], [24]. For example, the extraction of 34 fiducial points was proposed by Zhang et al. [5]. The displacement of these facial feature points between the current and previous frames is monitored to determine the facial movements in an image sequence. Valstar, Patras and Pantic demonstrated the effectiveness of tracked facial points and the detection of action units (AUs) in facial expressions [4]. Valstar and Pantic discussed the selection of most informative spatiotemporal features using AdaBoost and argued that the system could automatically track facial points and AUs. However, accurately detecting the facial points and tracking them is more challenging under noisy conditions.

Appearance based methods have also been used extensively to estimate the physical appearance of an image. To extract the appearance changes in facial images, holistic spatial analysis such as Gabor wavelet analysis [9], principal component analysis (PCA) [26], independent component analysis (ICA) [27] and linear discriminant analysis (LDA) [28] have been utilized frequently. Amin and Yan discussed the characteristics of multi-scale and multi-orientation Gabor filters for face recognition [8]. Donato, Bartlett, Hager, Ekman and Sejnowski employed PCA, ICA, LDA and Gabor wavelets to recognize the facial actions and suggested that the Gabor wavelets achieved the best recognition along with ICA [6]. Although the above-mentioned approaches achieved satisfactory results, there are a large number of Gabor features to be employed in classification. In this paper, we consider a small number of actual features by incorporating coherent patterns with a subset of samples and a subset of features. The selected features retain their original physical meanings and they are closely related to the samples that have specific characteristics.

2.2 Feature Selection

Feature selection for refining the feature representation and finding the semantic interpretation has also been employed in facial expression recognition. Feature selection can be performed by selecting certain informative spatial regions based on boosting learning. For example, LBP features can be extracted from the empirically weighted and equally divided small regions using Adaboost [17], [49], [51].

Recently, a work presented in [50] proposed stepwise linear discriminant analysis (SWLDA) for feature extraction. A set of localized features from face regions were selected during the training process and finally, hierarchical hidden conditional random fields were used for classification. Ligang et al. introduced the patch-based Gabor features that represent the salient regions of a face. They used Adaboost to evaluate and select image patches for each expression [52]. However, the learned patches from the same emotion are not consistent if different datasets are used. Happy et al. used facial landmarks to locate and evaluate salient regions and then computed the LBP features from these regions [51]. PCA-LDA is applied before classification.

In the above mentioned strategies, feature extraction from selected regions serves the purpose of reducing the identity bias. However, there is still a need to further transform the features to a new low dimensional feature space. In our method, rather than extracting features from predefined regions, we extract multiscale and multi-orientation Gabor wavelet features from the entire face and then select the relevant features. This effectively reduces the redundant information and noise and yield better accuracy as compared to existing approaches discussed above.

The recent success of deep learning based methods in different computer vision fields including FER has demonstrated encouraging performance. An Au-inspired feature learning framework was proposed in [16], [54], [55] to learn local textural patterns using a convolution layer on the apex expression frames. The learned features showed a strong descriptive power and physiological resemblance with the face AUs that encode the expressions. However, in our work using the proposed co-clustering discriminative feature learning, we are able to obtain local spatial regions near the AUs.

2.3 Classification

Several techniques have been proposed in the literature to accomplish the task of facial expression classification. They include support vector machine [7], nearest neighbours, neural network [3], [4], rule-based classifiers [20], [23], [24], and Bayesian network [35]. A comparison of SVM, AdaBoost, and LDA for facial expression recognition can be found in [7]. SVM is amongst more successful and effective learning method in FER [17], [49], [51], [52]. Therefore, we employ SVM for the classification task in our experiments. K-nearest neighbor (K-NN) classifier is an instance-based learning approach that uses small neighborhoods in the attribute space to predict the class label. These predictions can be significantly skewed by redundant attributes. However, after feature selection, the nearest neighbor approach performs better because of less noisy and refined features [18]. Therefore, we also adopt K-NN as an alternative classifier in our experiments.

2.4 Co-Clustering

To our knowledge, the idea of introducing co-clustering to select and localize features in appearance based feature extraction for facial expression recognition has never been attempted before. The process produces a spare set of most discriminant features for classification and reveals the relevant regions of these features.

Co-clustering methods were well-known for exploring the simultaneous row and column association for gene expression data analysis. A comprehensive survey of co-clustering (often called biclustering for 2D data) algorithms can be found in a review by Madeira and Oliveira [37]. Busygin et al. [38] discussed the applications of co-clustering in data mining for clustering words and documents simultaneously.

Noise in the data matrix remains a prime problem in co-clustering approaches. An algorithm based on exhaustive search, called large average submatrices (LAS), was proposed to overcome the influence of noise and discover overlapping co-clusters in the data matrix [36]. LAS is built on a heuristic randomized search to discover a co-cluster that maximizes the weight score on the residual matrix, which is obtained by subtracting identified co-cluster in successive iterations. An analysis of large-scale microarray data using co-clustering is presented by Zhao et al. [13]. A recent study on exploring multidimensional co-clusters using hyperplane detection in singular vector spaces is presented in [15]. In these methods, matrix factorizations, such as SVD and non-negative matrix factorization, have been explored to discover co-clusters in high dimensional data. The eigen basis methods allow a better representation of data using a smaller number of variables. The major difference in these factorizations lies in the sparse non-negativity constraints imposed on both dimensions of the data and therefore can reveal a more localized representation of patterns.

Nevertheless, all the existing approaches aim for discovering functionally related genes sets under several experiment conditions. However, in these approaches, no attention has been made to address class discriminant feature selection for classification. In this paper, we propose co-clustering based feature search strategy to effectively utilize the coherent structures from co-clusters. We derive a mechanism to select features, specifically in FER, where facial expression in one direction and the Gabor features in the other direction. The necessity of this mechanism lies in the fact that not all the co-clusters are useful and noisy ones within a class must be filtered out.

3 Feature Selection and Facial Expression Recognition System

In this section, we describe the general framework of our facial expression recognition (FER) system based on co-clustering as shown in Fig. 2. The system consists of three modules: Pre-processing, Facial feature extraction, and Recognition. Images are first pre-processed to normalize the face geometry, then features are extracted using a bank of Gabor filters. Co-clustering is then performed to produce the discriminant feature subset. The final stage contains a classifier. A multi-class SVM is applied to classify the prototype expression. The components of the FER system are discussed in the following sections.

3.1 Pre-Processing

To align the facial features, an image normalization is necessary. We first detect the face using Viola's face detector [42]

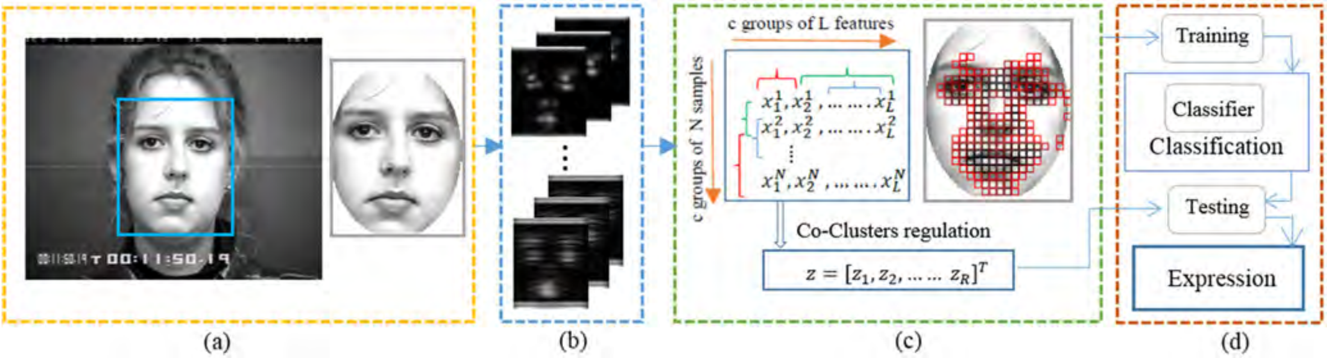
(a)
(b)
c groups of L features
c groups of N samples
$x_1^1, x_2^1, \ldots\ldots x_L^1$
$x_1^2, x_2^2, \ldots\ldots x_L^2$
$x_1^N, x_2^N, \ldots\ldots x_L^N$
Co-Clusters regulation
$z = [z_1, z_2, \ldots\ldots z_R]^T$
(c)
Training
Classifier
Classification
Testing
Expression
(d)

Fig. 2. The flow diagram of our method. (a) Pre-processing, (b) Gabor wavelet feature extraction, (c) co-clustering based feature selection and localization, and (d) classification.

and then normalize the images based on the eye positions to deal with the head rotation and imperfect localization. Cropping is done at a later stage using the ellipse face mask. Furthermore, despite the presence of illumination variations in images, no intensity normalization was performed because the Gabor wavelets are grayscale invariant.

3.2 Gabor Wavelets for Image Representation

Wavelet filters are capable of decomposing an image into appropriate texture features. Due to their relevance to the human visual system, multi-channel filtering has gained much attention in computer vision for recognition tasks.

3.2.1 *Gabor Function*

The Gabor function in the spatial domain represents a Gaussian-shaped envelop modulated by a complex sinusoidal signal [44], [45]

$$\mathrm{g}(\mathrm{x},\mathrm{y})=\frac{1}{2\pi\delta_{\mathrm{x}}\delta_{\mathrm{y}}}\exp\left\{-\frac{1}{2}\left[\left(\frac{\mathrm{x}}{\delta_{\mathrm{x}}}\right)^2+\left(\frac{\mathrm{y}}{\delta_{\mathrm{y}}}\right)^2\right]+\mathrm{i}(\mathrm{ux}+\mathrm{vy})\right\}. \tag{1}$$

In the frequency domain, the Gabor function can be viewed as a 2-D Gaussian function or simply a band-pass filter

$$\hat{\mathrm{g}}(\mathrm{w}_{\mathrm{x}},\mathrm{w}_{\mathrm{y}})=\exp\left\{-2\pi^2\left[\delta_{\mathrm{x}}^2(\mathrm{w}_{\mathrm{x}}-\mathrm{u})^2+\delta_{\mathrm{y}}^2\left(\mathrm{w}_{\mathrm{y}}-\mathrm{v}\right)^2\right]\right\}. \tag{2}$$

The Gabor function is capable of delivering maximum possible resolution in the frequency domain, and vice versa [47]. The Gabor transformation makes use of a set of functions with multiple scales and orientations as its basis functions. When used as a feature extractor, the Gabor wavelets preserve the spatial structure of an image while extracting the frequency contents of the image. The real and imaginary parts of a Gabor function can be represented as [9], [45]

$$\begin{aligned}
G_{\vec{k}}(\vec{r}) &= G_{\vec{k},+}(\vec{r})+iG_{\vec{k},-}(\vec{r})\\
G_{\vec{k},+}(\vec{r}) &= \frac{k^2}{\delta^2}\exp\left(\frac{k^2\|r-r_o\|^2}{-2\delta^2}\right)\cos\left[\vec{k}(\vec{r}-\vec{r_o})\right]\\
G_{\vec{k},-}(\vec{r}) &= \frac{k^2}{\delta^2}\exp\left(\frac{k^2\|r-r_o\|^2}{-2\delta^2}\right)\sin\left[\vec{k}(\vec{r}-\vec{r_o})\right],
\end{aligned} \tag{3}$$

where, $\vec{k}=k\exp(j\theta_v)$ and k represents the scale and θ the rotation.

3.2.2 *Feature Extraction Using a Bank of Gabor Filters*

For feature extraction, different parameters of the Gabor function are responsible for representation of complementary information. Consequently, we use a bank of Gabor filters, and employ 5 scales $(k=\frac{\pi}{2\sqrt{2^r}}, r=1,2,3\ldots,5)$ and 8 orientations $(\theta_v=\frac{u\pi}{8}, u=0,1,2\ldots,7)$, with $\delta=\pi$. The Gabor wavelet based image representation is obtained by convolving the Gabor filter bank with the facial image

$$R_{\vec{k},\pm}(\vec{r}_0)=\int G_{\vec{k},\pm}(\vec{r}_0,\vec{r})I(\vec{r})d\vec{r} \tag{4}$$

$$R_{\vec{k}}=\sqrt{R^2_{\vec{k},+}+R^2_{\vec{k},-}}. \tag{5}$$

For an input image, we compute 40 Gabor responses (for 5 scales and 8 orientations) and generate the feature vector based on the amplitude of the filter output. Down-sampling is usually required due to the presence of high spatial correlation among the neighboring pixels.

For the JAFFE database, the image size is 150×110 after pre-processing. Down-sampling rows and columns with the factor of 12, we obtain a vector of length 130 ($\lceil 150/12\rceil\times\lceil 110/12\rceil=13\times10$, where, $\lceil \mathrm{x}\rceil$ is the least integer greater than or equal to x). Similarly, in CK+ and MMI databases with the image size of 280×230, after downsamling by the factor of 12, concatenation will produce a vector of length 480 ($\lceil 280/12\rceil\times\lceil 230/12\rceil=24\times20$). This means the resulting feature vector from the bank of filters for the JAFFE database will have a total length of $130\times40=5200$ and $480\times40=19200$ for CK+ and MMI. This high dimensionality issue can have an undesirable impact on the effectiveness of the learning algorithm [21].

Feature selection, therefore, can be beneficial under the high dimension and low sample space (HDLSS) scenario, where the dimensionality of the feature vector is substantially higher as compared with the number of available training images. The problem of HDLSS makes it challenging for the conventional multivariate analysis [30]. In most multivariate analysis methods, the preliminary step is usually to 'sphere the data', which is obtained by the product of root inverse of the covariance matrix and the data matrix. However, for HDLSS data, this inverse does not exist when the covariance matrix does not have the full rank. Under these conditions, it might be reasonable to consider that most features do not contribute to a large extent or do not distinguish among several expression classes, and can be regarded as

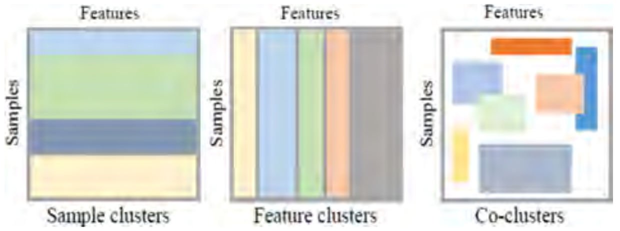

Fig. 3. An illustration of the conventional clustering (left and middle diagrams) and co-clustering (right diagram).

noise. Co-clustering is powerful for solving the HDLSS problem in facial expression recognition. As we will see in the following sections, only a small number of Gabor wavelet features around the eyes and mouth are needed for facial expression recognition. Based on co-clustering, we can identify these features and filter out others, which are irrelevant or can be considered as noise for the recognition purpose.

3.3 Co-Clustering Based Feature Selection (CCFS) Method

Co-clustering is an unsupervised learning procedure that targets the association of rows and columns in the feature matrix, yielding distinct “checkerboard” patterns in the data matrix. Recently, co-clustering methods have gained much popularity and found many applications in data mining and biological studies [31], [32], [33].

Classical clustering techniques perform classification in one direction only to deal with the similarities of samples (in rows) as reflected by their features (in all columns) of a matrix. They typically assume that samples in a particular cluster share exactly the same properties over all available features. Therefore, conventional clustering methods reflect the global patterns of samples and ignore the local patterns among samples and features [13]. In reality, different samples in a data matrix may have properties reflected by different features. The co-clustering approaches identify the local patterns in the data matrix that are apparently not visible. In contrast to one-dimensional clustering where disjoint clusters are formed that covers all available features, co-clustering performs clustering in two directions simultaneously and produces co-clusters that may overlap and only represents a part of the matrix as illustrated in Fig. 3. These local patterns reveal a joint similarity in samples among specified features which can be useful in distinguishing one class from others. The following sections provide an analysis of CCFS, whereas the procedure for CCFS is summarized in Algorithm 1.

3.3.1 SVD Connection

To detect co-clusters from a data matrix, a number of co-clustering approaches have been developed based on matrix factorization techniques, including singular vector decomposition (SVD) [15] and its higher order forms [13], [25], [47]. This paper focuses on SVD to decompose the facial expression data matrix. SVD has been explored to detect arbitrary combinations of the key features as co-clusters. In order to extract useful coherent patterns from the data and overcome the noise influence, only the first several singular vectors corresponding to the largest singular values are selected. In this way, SVD can produce more localized feature representations of both expression samples and features.

Algorithm 1. CCFS for Feature Selection

```
Input: Training feature matrix $A_{m\times n} = [a_1, a_2, \dots a_n]$
  $Q$ the number of co-clusters , $D$ singular vectors.
Output: $g_j = [d_{1j}\ d_{2j} \dots d_{Rj}]$, $R < \mathrm{N}$
  Selected features $g_j$ and indices $\mathrm{f}_R$
1: Apply SVD to the standardized feature matrix and select top
   $D$ singular vectors.
2: Perform co-clustering by executing independent iterative
   clustering on row space and column space following the
   formulations: Eqs. (13), (14), (15)
3: Select candidate co-clusters based on class membership
   using Eq. (16)
4: Return feature subset indices based on non-inclusive
   information in overlapped co-clusters obtained above using
   Eqs. (17) and (18)
```

Let us consider a data set of m samples and n variables, forming a rectangular matrix $A = (a_{ij})_{m\times n}$, where a_{ij} represents the ith facial expression sample and the jth feature. First, we normalize the features for $i = 1, \dots, n$ to have zero mean and unit variance as

$$\hat{a} = \frac{a_i - \hat{a}_i}{s_i}(i = 1, \dots, n). \tag{6}$$

$$\hat{A}_{m\times n} = [\hat{a}_1, \hat{a}_2, \dots \hat{a}_n]. \tag{7}$$

The SVD of data matrix $\hat{A}$ can be expressed as

$$\hat{A} = U\Lambda V^T = \sum_{i=1}^{r} \lambda_i u_i v_i^T, \tag{8}$$

where, $U = [u_1, u_2, \dots u_r]$ and $V = [v_1, v_2, \dots v_r]$ represents the orthogonal left and right singular vectors respectively, r is the rank of $\hat{A}$, and $\Lambda = \mathrm{diag}(\lambda_1, \lambda_2, \dots \lambda_r)$ is the diagonal matrix containing ordered singular values $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_r > 0$. For the largest λ_k value, $\lambda_i u_i v_i^T$ represents an SVD layer responsible for the most significant information related to data matrix. Representation of the data matrix in an ideal case with checkboard patterns $A_1, \dots, A_k$ takes the form

$$\hat{A} = \begin{bmatrix} A_1 & 0 & \cdots & 0 \\ 0 & A_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & A_k \end{bmatrix}. \tag{9}$$

Fig. 4 shows the 3D visualization of co-clusters in singular vector spaces. The actual data is prepared from a matrix initially containing zero values. Co-clusters with large values are added to form checkboard patterns in the matrix. The noisy data are generated by first randomly shuffling the actual arrangement of checkboard patterns and then adding normally distributed noise of standard deviation 0.5. With a rank 3 approximation of the data matrix, we can visualize the co-clusters in the 3D space of three right singular vectors, and the space of three left singular vectors, as shown in the two diagrams in Fig. 4. The five clusters in the singular vector spaces correspond to the four co-clusters in the checkboard patterns plus the background of the data matrix.

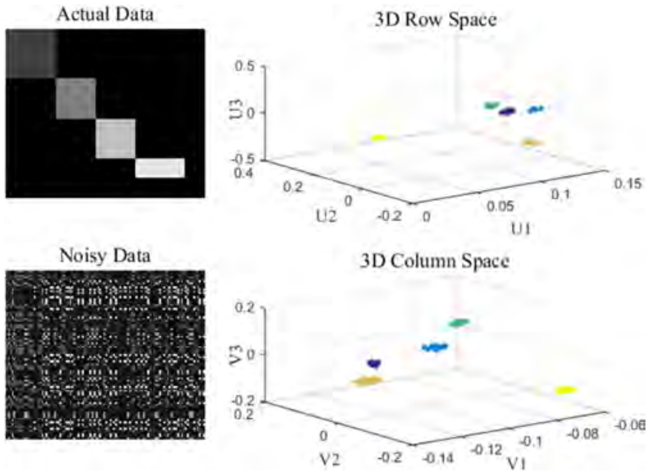

Fig. 4. An example showing the row and column association in co-clusters. Rows and columns in the synthetic checkboard patterns (upper left) are reordered randomly and then the data matrix is added with noise (lower left). Five clusters in the singular vector spaces (two diagrams on the right) correspond to the four co-clusters in the checkboard patterns plus the background of the data matrix.

3.3.2 Co-Clustering

In the facial expression feature matrix, a set of similar facial expression samples are expected to have a correlation in terms of a set of features, as shown in Fig. 1. We explore these submatrices in the dominant singular vectors space. In Eq. (8), we denote the row space and column space as S_R and S_C respectively as

$$S_R = [u_1, u_2, \ldots u_d]^T = [x_1, x_2, \ldots x_m] \tag{10}$$

$$S_C = [v_1, v_2, \ldots . v_d]^T = [y_1, y_2, \ldots y_n]. \tag{11}$$

For a given number of targeted co-clusters Q, we compute the smallest integer $\tilde{Q}$ not less than $\sqrt{Q}$, where $\tilde{Q} = \sqrt{Q}$ and divide the elements in S_R randomly in $\tilde{Q}$ groups as: $G^0 = \{g^0_1, \ldots g^0_{\tilde{Q}}\}$

Sample assignment to a group can be computed according to the probability B^l_{ij} in Eq. (12). That is, for a given sample x_j belonging to group g^l_i, $l = 0, \ldots .\ L$, the probability is caulculated as

$$\mathrm{B^l_{ij}} = \frac{1}{2\pi\sqrt{\det\left(\mathrm{C^l_i}\right)}} \mathrm{e}^{-\frac{1}{2}\left(\mathrm{x_j}-\mathrm{m^l_i}\right)' * \left(\mathrm{C^l_i}\right)^{-1} * \left(\mathrm{x_j}-\mathrm{m^l_i}\right)}, \tag{12}$$

where, $\mathrm{i} \in \{1, \ldots, \hat{\mathrm{Q}}\}$, $\mathrm{j} \in \{1, \ldots, \mathrm{m}\}$, m^l_i is the mean and C^l_i is the covariance matrix of samples x_j in group g^l_1. Sample assignment is updated to g^{l+1}_1 as

$$\mathrm{B^l_{\hat{i}\hat{j}}} = \max_{\mathrm{i} \in \{1,\ldots,\hat{\mathrm{Q}}\}} \mathrm{B^l_{ij}}. \tag{13}$$

For each iteration, based on the assessment function, we select the G group corresponding to the minimum value of D^l. The assessment function for each iteration can be computed as

$$D^l = \sum_{i \in \{1,\ldots,\hat{Q}\}} \|C^l_i\|, \tag{14}$$

where, $\|C^l_j\|$ represents the Frobenius norm of the covariance matrix C^l_i.

The same process is then repeated for S_C. Row information from S_R and column information from S_C are combined in $\tilde{Q}$ groups to form $\tilde{Q}^2$ submatrices.

In order to recognize the scaling patterns of the identified submatrices, we use the following evaluation criterion that computes the mean squared residual score [15] to select the Q co-clusters among available $\tilde{Q}^2$ sub-matrices

$$H(I, J) = \frac{1}{|I||J|} \sum_{i \in I, j \in J} \left(a_{ij} - \bar{a}_{Ij} - \bar{a}_{iJ} + \bar{A}_{IJ}\right)^2, \tag{15}$$

where, A_{IJ} is the sub-matrix obtained from the data matrix A, $I = \{i_1, i_2, \cdots i_s\}$ and $J = \{j_1, j_2, \cdots j_t\}$ are associated subsets of rows and columns respectively, $\bar{a}_{iJ}$ and $\bar{a}_{Ij}$ are the means of the ith row and jth column respectively, and $\bar{A}_{IJ}$ is the mean of sub-matrix IJ.

3.3.3 Feature Selection

In terms of local patterns, the co-clustering method exhibits an overwhelming advantage over hard clustering. In co-clustering, we are able to identify a small subset of sample and feature correspondences, where several co-clusters can partially overlap. For instance, a co-cluster, (I, J) is a $(s \times t)$ submatrix of the original $(m \times n)$ matrix, having the column indices, $J = \{j_1, j_2, \cdots j_t\}$. By looking at expression samples in one direction and features in the other direction, the relevancy of features can be determined. The procedure looks for a set of features that distinguish a subset of different class samples from each other. In order to achieve this, we compute the probability of each class in co-clusters as

$$P^c_Q = \frac{\sum_c E^c_Q}{\left|\mathrm{I_Q}\right|} \in [0\,1], \tag{16}$$

where E^c_Q is the number of elements in co-cluster $\mathrm{I_Q}$ that belongs to class c. A higher probability indicates that the related co-cluster contains sufficient representation from a class. Candidate co-clusters that have the maximum class sample probabilities i.e $\max(P_s)$, are retained for feature selection, where $P_s \subset P^c_Q$ and $s\ \epsilon\ [e, t]$.

In between e and t of co-clusters J, there exist common indices of columns, given by

$$f_e = \{j_{\mathrm{e}} \cap j_{\mathrm{e+1}} \ldots j_t\}, e \leqq t \tag{17}$$

$$f_R = \{f_{\mathrm{L}} - f_e\}. \tag{18}$$

Indices of selected features are given by the compliment of the original feature indices f_{L} of length L and f_c.

Thus, we can form the feature matrix of reduced dimensionality

$$g_j = \left[d_{1\mathrm{j}}, d_{2\mathrm{j}} \ldots d_{\mathrm{Rj}}\right]^T, j = 1, 2 \ldots N. \tag{19}$$

3.4 Classification Using Multiclass SVMs

For facial expression recognition, the final task of classification is performed with SVM, which is a binary discriminant classifier built based on the structural risk minimization principle that creates maximum margin hyperplane among two classes [14]. In our experiments, we used the multiclass SVMs and the one against all strategy. Detailed illustration about multiclass SVMs and their formulation can be found in [14], [22], [29]. Here we briefly present the optimization problem of multiclass SVMs followed in our experiments.

Given the training data $(g_1, l_1), \ldots, (g_{\mathrm{N}}, l_N)$ where, $g_j \in \Re^R$ is a reduced Gabor feature vector and $l_j \in \{1, \ldots, 7\}$ is the corresponding expression label of a feature vector. Multiclass SVMs targets only one optimization problem [29] but constructs seven class rules. Incorporating the $k\mathrm{th}$ function $\boldsymbol{w}_k^T \boldsymbol{\phi}(g_j) + b_k$ to partition training vectors of class k from the rest of feature vectors, we minimize the objective function

$$\min_{w,b,\xi} \frac{1}{2} \sum_{k=1}^{7} \boldsymbol{w}_k^T \boldsymbol{w}_k + \sum_{j=1}^{N} \sum_{k \neq l_j} \xi_j^k, \tag{20}$$

subject to the constraints

$$\begin{aligned} &\boldsymbol{w}_{l_j}^T \boldsymbol{\phi}(g_j) + b_{l_j} \geq \boldsymbol{w}_k^T \boldsymbol{\phi}(g_j) + b_k + 2 - \xi_j^k \\ &\xi_j^k \geq 0, j = 1, \ldots, N, k\{1, \ldots, 7\} \backslash l_j, \end{aligned} \tag{21}$$

where, ϕ represents the mapping function, C penalizes the training errors, $b = [b_1 \ldots b_7]^T$ is the bias vector, ξ is a slack variable, and $\xi = [\xi_1^1, \ldots, \xi_i^k, \ldots, \xi_N^6]^T$. The decision function is given by

$$h(g) = \underset{k=1,\ldots,7}{\arg\max} \big(\boldsymbol{w}_k^T \boldsymbol{\phi}(g_j) + b_k\big). \tag{22}$$

After training the seven class SVMs, a new test Gabor feature vector is classified using the equation above to recognize the facial expression.

4 Experiments

This section provides the detailed analysis and comparison of facial expression recognition based on the proposed feature selection strategy.

4.1 Experiment Setup

In our experiments, we focused on three widely used databases: JAFFE (Japanese female facial expression database) [9], Cohn–Kanade (CK+) [43] database and MMI [49]. JAFFE database is commonly used in FER systems, comprising of 213 grayscale images of 213 × 213 resolution from 10 Japanese female participants for which two to four images are obtained for a single facial expression. All 213 images are used in our experiments.

The Cohn-Kanade database includes a diversity of participants and is currently considered being one of the most comprehensive human face image databases. The database comprises of 100 university students including both male (35 percent) and female (65 percent) aged from 18 to 30 years from different origins. For a single prototype emotion, there is a series of image sequence starting from neutral to a peak intensity of a target emotion. In our experiments, we selected one neutral frame for a single subject and 3 peak frames for the rest of the prototype emotions. For 83 subjects, each subject had at least one prototype emotion. This resulted in 543 images in total (78 Neutral, 78 Happiness, 78 Disgust, 78 Surprise, 69 Fear and 78 Sadness). The MMI database consists of 30 subjects of different sexes and age from 19 to 62 years of age with the multiethnic background. In this database, 205 out of 213 sequences have the frontal view with emotion labels. We selected 90 sequences having 1-6 basic emotions per person. Similar to CK+, we selected one neutral frame and 3 peak frames from each sequence, yielding 381 images in total with approximately balanced class size.

We adopted 10-fold cross validation strategy, in order to evaluate the generalization performance of FER system. More specifically, all the subjects were divided into ten folds of nearly equal size. Nine folds were used to train the classifier using the available emotion labels and the remaining fold was left for testing. This process was repeated 10 times so that each group for training was also used for testing. The performance of the classifier is expressed in terms of average recognition rate and the F-score. Confusion matrices [9] are also computed by providing the true class emotions l_{tc} (columns) and recognized class emotions l_{rc} (rows).

To demonstrate the effectiveness of the proposed CCFS algorithm, we compared following state of the art feature selection algorithms that rank features based on their significance. In addition, another co-clustering algorithm was also considered, to conduct feature selection.

- LS-FS (Laplacian score) proposed in [39], which selects features that are consistent with Laplacian Gaussian matrix and best aligns with the manifold structure of the data.
- LLC-FS (Local learning-based Clustering) [40], in which feature selection is carried by using local learning based clustering and training the local regression model with the points in each neighborhood.
- Inf-FS (Infinite feature selection) [41] is a graph based feature selection method, where each feature corresponds to a node in a graph and path represents the feature selection. Centrality score is assigned to the important feature by considering other feature subsets as a path on graph.
- LAS (Large average submatrices) [36] is a co-clustering method, which operates on heuristic (nonexact) and randomized search driven by significance score-based function that tradeoff between average value and size of submatrix.

We used the following parameter values: the number of singular vectors: $D = 10$, the number of submatrices: $N = 50$, and the number of groups: $\tilde{Q} = 5$. Parameter e in Eq. (18) represents the size of a feature subset. We choose several values in our experiments to produce feature subsets of different sizes and are represented as CCFS(C), where the subscript is used to differentiate feature subsets and the value in subscript (C) indicates the number of co-clusters used to produce this distinct feature subset. The recognition rate corresponding to each feature subset is shown in Fig. 5. It is evident that the recognition rate increases when the size of feature set increases, and reaches up to an optimal value, increasing further a number of selected features may add extra noise and degrades the performance. In this case, low sample-to-feature-ratio also affect the performance of learning algorithm, whereas an adequate feature subset selection not only improves the scalability but also defies the identity bias in FER [12].

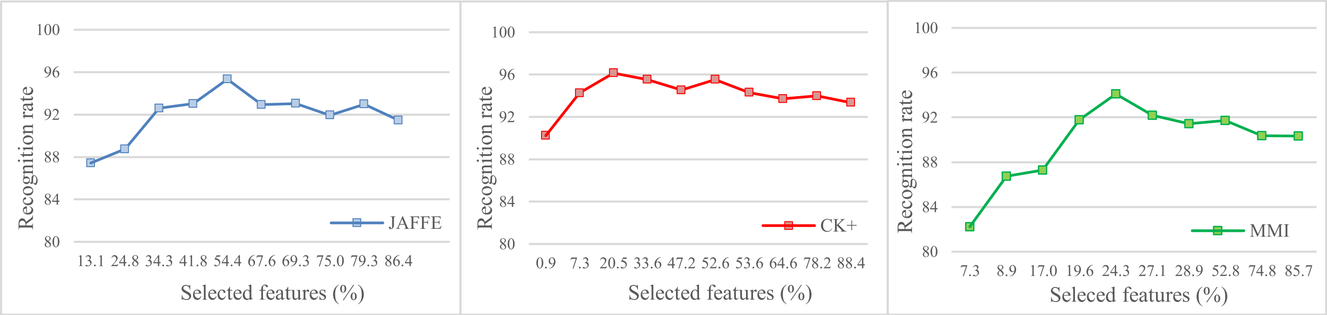

Fig. 5. Average recognition rate obtained by proposed CCFS using SVM with feature subsets of varying sizes on Jaffe, CK+, and MMI.

4.2 CCFS Representation

The feature selection strategy, where a discriminant subset of a feature is chosen to characterize the high dimensional feature matrix, has a significant effect, not only on the recognition rate but also on the computational complexity. However, the feature subset must have adequate information, such that an appropriate recognition of a test image can be carried out. As discussed in Section 3.3, the CCFS explores the block structures in a feature matrix. Based on the class membership in co-clusters, we make the feature selection. The class probabilities in the co-clusters for CK+ are shown in Fig. 9. Some of the co-clusters do not have any participation in any classes, so these co-clusters are neglected. As evident in Fig. 9, co-clusters 0 to 5 do not have any occurrence in classes 1 and 2. On the contrary, the co-clusters that have the maximum class probabilities and also have the maximum number of samples are first selected due to their low mean-square residue and high sample–variance.

We refer to Fig. 6 to understand how the individual Gabor filters are accountable for selecting specific regions of the face. First, it can be observed that feature selection in each filter corresponds to those regions that have the maximum response at particular scales and orientations. Second, not all the filters are equally important. Some are totally redundant in terms of discriminant power. Therefore, many features form different response filters in a region can be selected and less informative regions can be discarded. By combining all selected features corresponding to several spatial locations from different filters, we reduce the possibility of missing features from salient face points, and produce compact local features.

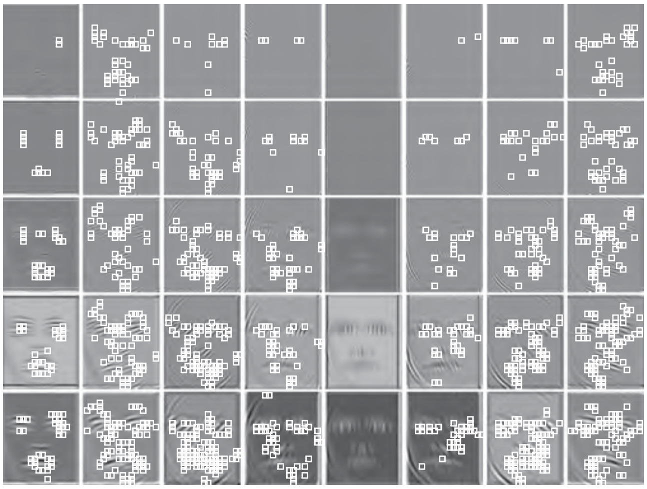

Fig. 6. Visualization of CCFS mapping on the real part of filters in the local Gabor filter bank. Each row corresponds to one filter with varying scales from top to bottom, whereas the column represents the varying orientation from left to right. The spatial patterns contributed by each filter resembles the informative regions of the face.

Fig. 7a demonstrates the spatial representation of few top selected feature points on the JAFFE dataset. Among 40 Gabor filtered images, the selected features are arranged in the order of maximum occurrence as shown in Table 1. Feature points at locations 1 and 2, which are found in the mouth region has the highest occurrence and were found repeatedly in 39 and 37 Gabor images respectively. Feature points 3, 4, 5, 6, and 7 are found at the eye locations with 36 and 35 occurrences respectively. Similarly, Figs. 7b and 7c show the distribution of feature points from the CK+ database images. Fig. 7b is chosen to illustrate the distribution of the minimum number of selected features, i.e., 171 out of total 19200 features. Only 97 features can be plotted due to different scales and orientations at the same location. Figs. 7c represents the distribution of the selected features that achieved the highest recognition rate in our experiments. Fig. 7d shows the locations of 9 percent features selected from the MMI dataset, and Fig. 7e the selected features that produce the highest recognition rate. The results corresponding to each of these feature representations are discussed in a later section. It is evident from these results that only the salient regions of face parts are spotted that account for the movement of facial expression muscles.

Apart from CCFS, we also conducted experiments on other feature selection methods discussed above. The feature locations obtained using these methods with similar numbers of top ranked features in each of the databases are shown in Fig. 8. For CK+ and MMI, the results of feature selection are marginally acceptable in terms of salient regions. However, it can be noted that noisy features such as boundary regions are also selected, due to image illumination variations present at cropping regions.

4.3 Facial Expression Recognition Rates

We compared the performance of different feature selection algorithms. Table 2 compares the recognition rates from the CK+ dataset. The number of extracted features is 19200, which is relatively high as compared to the number of available sample images, 535. With this high dimensional feature matrix, we obtained a recognition accuracy of 76.32 percent using K-NN (K = 1) and 87.37 percent using SVM (rbf kernel). We evaluated two feature sets named as CCFS (C7) and CCFS(C12) having different numbers of features

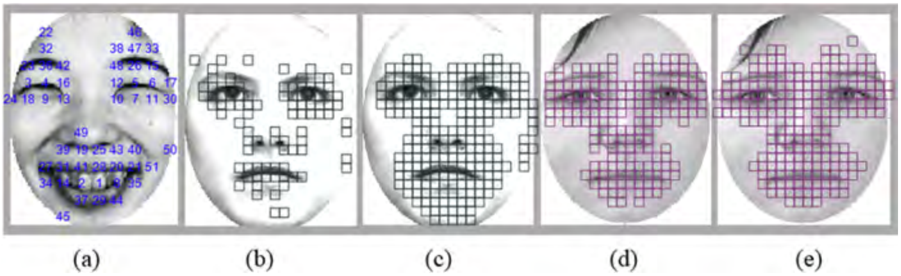

Fig. 7. Distribution of the feature points selected using CCFS on a) Jaffe database, (b),(c) on CK+ database, and (d),(e) on MMI database.

which are obtained from proposed co-clustering feature selection method. The number in the parenthesis denotes the number of co-clusters used in the feature selection process. Here CCFS(C12) represents the feature subset that achieved best recognition rate whereas CCFS(C7) is chosen to demonstrate the performance on the minimum number of features.

It is evident from the Table 2 that the proposed method using CCFS(C12) achieved the best performance in terms of average recognition rate followed by clustering based method, LLC-FS and the filter method Inf-FS. The visualization of features corresponding to CCFS(C12) is presented in Fig. 7c, from which we can observe that that the CCFS method captured only those regions that are responsible for expressions. While using these 3891 features, which are just 20 percent of the original features, we obtained a recognition rate of 85.40 percent with K-NN. For the same set of features, SVM provided a recognition rate of 96.05 percent, which is the highest accuracy in our experiments with an Fscore of 96.14 percent. The reason for this increase in accuracy is the removal of the effect of noise by selecting only the regions that are accountable for facial expressions.

An interesting result with CCFS(C12) is obtained when we used just 171 features which is only 0.9 percent of the original features. With such a small number of features, a recognition rate of 90.23 percent was achieved, which is still better than the accuracy obtained with the full set of features. The distribution of these features is shown in Fig. 7b. This indicates that our feature selection strategy based on co-clustering preserves the discriminating strength of features among classes.

The confusion matrix for the best predicted results with CCFS(C7) and SVM is given in Table 5, which shows that two expressions, i.e., Happy and Surprise, are recognized

TABLE 1
Occurrence of Fiducial Points in Gabor Images

Mark	Outcome	Mark	Outcome	Mark	Outcome
1	39	12-14	32	32	25
2	37	15-21	30	33-35	24
3	36	22-25	29	36-37	23
4-8	35	26-28	28	38-41	22
9	34	29	27	42-45	21
10-11	33	30-31	26	46-51	20

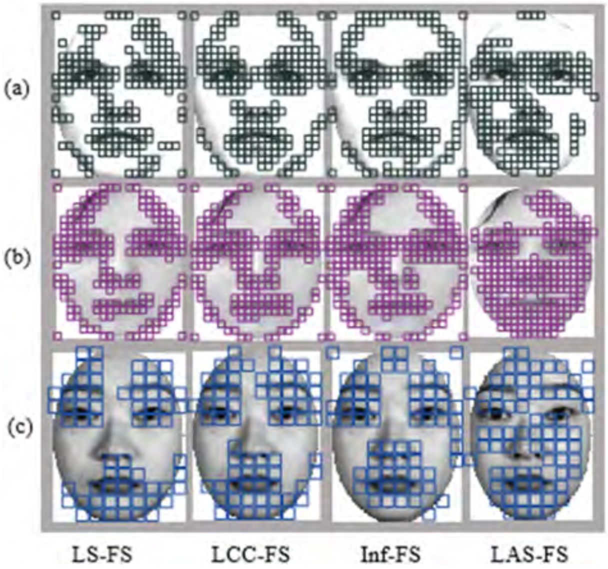

Fig. 8. Distribution of selected feature points using LS-FS (Laplacian score feature selection), LLC-FS (Local learning-based Clustering feature selection), Inf-FS (Infinite feature selection) and LAS (Large average submatrices co-clustering feature selection) on (a) CK+, (b) MMI and (c) Jaffe datasets.

TABLE 2
Recognition Performance Comparison on CK+

Database	Method	Recognition rate	
		KNN	SVM
CK+	LAS-FS	74.44	85.20
	LLC-FS	83.27	94.65
	LS-FS	79.98	88.43
	Inf-FS	80.56	91.08
	CCFS (C7)	85.40	**96.05**
	CCFS (C12)	82.15	90.23
	AF	76.32	87.37

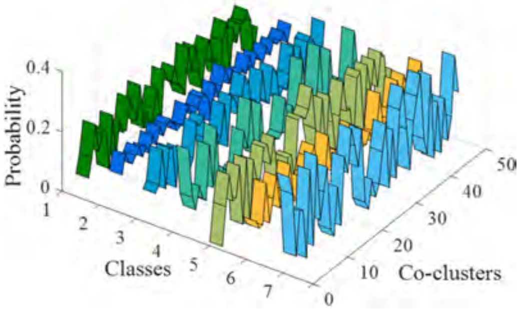

Fig. 9. Probabilities of samples from each class in co-clusters (membership) obtained from CK+ dataset. Co-clusters with high sample variance form each class are shown with larger peaks. The process is designed to incorporate the candidate biclusters for feature selection.

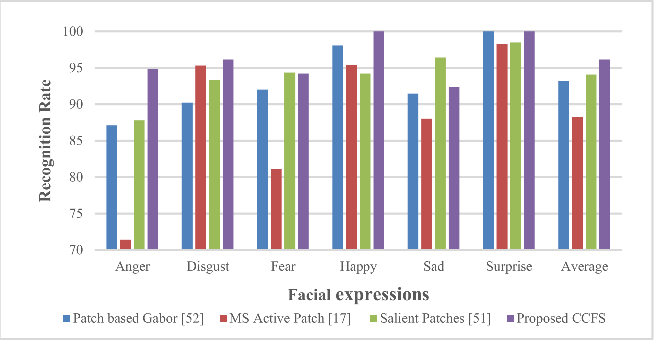

Fig. 10. Recognition rate comparison of six facial expressions on CK+ with closely related work. The rightmost four vertical lines represent the average recognition rate for all six facial expressions.

TABLE 3
Recognition Performance Comparison on JAFFE

Database	Method	Recognition rate	
		KNN	SVM
JAFFE	LAS-FS	75.38	87.33
	LLC-FS	86.54	95.11
	LS-FS	75.96	85.87
	Inf-FS	84.55	93.08
	CCFS (C5)	87.42	**95.31**
	CCFS (C9)	83.30	88.74
	AF	77.32	88.39

TABLE 4
Recognition Performance Comparison on MMI

Database	Method	Recognition rate	
		KNN	SVM
MMI	LAS-FS	70.33	79.50
	LLC-FS	81.67	91.77
	LS-FS	69.19	78.45
	Inf-FS-FS	78.90	90.88
	CCFS (C7)	83.22	**94.09**
	CCFS (C10)	81.66	86.74
	AF	71.12	84.22

TABLE 5
Confusion Matrix of CCFS **(C7)** Using SVM on CK+

ltc/lrc	Average recognition rate = 96.05%						
	Ang.	*Disg.*	*Fea.*	*Hap.*	*Sad*	*Sur.*	*Neut.*
Ang.	**94.87**	0	0	0	1.28	0	3.85
Disg.	3.85	**96.15**	0	0	0	0	0
Fea.	0	0	**94.20**	0	4.35	0	1.45
Hap.	0	0	0	**100**	0	0	0
Sad	2.56	0	0	0	**92.31**	0	5.13
Sur.	0	0	0	0	0	**100**	0
Neut.	0	0	0	0	5.13	0	**94.87**

TABLE 6
Confusion Matrix of CCFS **(C5)** Using SVM on JAFFE

ltc/lrc	Average recognition rate = 95.31%						
	Ang.	*Disg.*	*Fea.*	*Hap.*	*Sad*	*Sur.*	*Neut.*
Ang.	**93.33**	0	0	3.33	3.33	0	0
Disg.	0	**96.77**	0	0	0	0	3.23
Fea.	0	3.33	**96.67**	0	0	0	0
Hap.	0	3.125	0	**93.75**	0	3.13	0
Sad	0	0	0	0	**93.33**	6.67	0
Sur.	0	0	0	0	0	**96.55**	3.45
Neut.	0	3.23	0	0	0	0	**96.77**

TABLE 7
Confusion Matrix of CCFS **(C7)** Using SVM on MMI

ltc/lrc	Average recognition rate = 94.09%						
	Ang.	*Disg.*	*Fea.*	*Hap.*	*Sad*	*Sur.*	*Neut.*
Ang.	**91.16**	0	0	3.33	5.00	0	0
Disg.	1.82	**98.18**	0	0	0	0	0
Fea.	0	0	**89.36**	0	0	6.64	4
Hap.	0	0	0	**92.16**	0	0	7.84
Sad	0	0	0	0	**98.00**	0	2
Sur.	0	0	5.17	0	0	**94.83**	0
Neut.	0	0	0	0	5	0	**95.00**

effectively. However, Sad is difficult to predict, as it confuses with Neutral, and this might be due to the fact that intensity of the Sad expression is not high on several subjects. Fig. 10 shows the comparison of emotion rates of various approaches on CK+ considering the 6-class problem. The proposed approach achieves the highest average recognition rate. For Surprise, patch-based Gabor approach and proposed CCFS achieved 100 percent. For Anger, Disgust and Happy, CCFS also performed better. However, for Sad, the performance was inferior to that of [51], due to the misclassification with Neutral, which is not considered in [51].

The results obtained on the seven-prototype-expression for JAFFE database are shown in Table 3. The highest recognition rate of 95.31 percent and Fscore of 95.33 percent are achieved using SVM on CCFS(C5). K-NN on the other hand achieved the maximum recognition rate of 87.42 percent on CCFS(C5), which is 10 percent increase as compared to AF. The confusion matrix for CCFS(C5) is shown in Table 6, in which Sad and Anger expressions achieved low rates, whereas Neutral and Disgust performed comparatively better. The results from LLC-FS and Inf-FS were generally better than AF. However, LAS-FS and LAS-FS were found to produce recognition rates lower than the baseline and this degradation is mainly due to avoiding key features from the salient regions, as evident in Fig. 8.

The results from MMI are presented in Table 4 in which the proposed CCFS algorithm consistently outperforms all other methods. The best recognition rate of 94.09 percent is obtained using CCFS(C7) with Fscore of 94.19 percent. The selected features are shown in Fig. 7e. Laplacian score based feature selection showed the worst feature ranking as the recognition rate was found to be the lowest. This might be due to the lower sample-to-feature ratio present in the MMI dataset, and this makes it difficult to preserve the local manifold structure. In our experiments, MMI achieved lower recognition performances as compared to CK+ and JAFFE, due to the fact that the subjects in MMI are occluded with glasses and mustaches. Also there is a variation within expressions among several subjects. The confusion matrix for MMI is given in Table 7. The expressions of Fear and Anger have lower recognition rates than others, whereas Disgust achieved the highest rate of 98 percent.

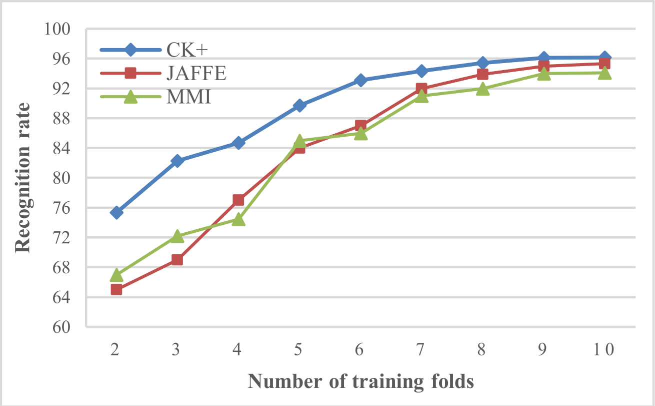

Fig. 11. The effect on recognition rate by increasing the value of k in k-fold on CK+, JAFFE and MMI databases.

4.3.1 Effect on Varying k in k-Fold

For all three databases, we partitioned the data using different values of k in k-fold cross validation to study the impact of changing the proportion of training and testing samples. Results in Fig. 11 show the accuracy of SVM on the best features selected from the three databases for different values of k. When half of the data is used for training and the remaining half for testing, we still achieved more than 75 percent for CK+ and 65 percent for both JAFFE and MMI. While increasing the number of training samples and subjects, there is an increase in accuracy and we found that after 7-folds the accuracy of all three databases approaches more than 90 percent. The recognition performance is better with the JAFFE database, which only has 12 subjects of the same gender, as compared to MMI. The MMI database also does not have large enough samples but has more than twice the subjects as compared to JAFFE, which makes the recognition task difficult.

The training time for CCFS varies for a particular database as the number of training samples in k-fold increases. We computed the execution time of CCFS with MATLAB on a Dual core Intel Pentium i7 CPU with 3.6 GHz & 8 GB memory. Using the parameter settings stated in Section 4.1, the CCFS implementation with $k = 2$ in k-fold, requires training time (in sec) of 562, 380 and 140 for CK+, MMI and Jaffe databases respectively. Whereas for higher values of k, e.g., $k = 10$, the computation time (in sec) decreases linearly to 344, 210 and 90 for CK+, MMI and Jaffe databases respectively.

4.3.2 Experiments on Cross-Database

The proposed CCFC was used to perform the inter-database experiments. As we pre-processed CK+ and MMI images in the same manner, it is possible to train and test on either of the databases. First, we used 537 available images from CK+ to train the SVM classifier using the CCFS learned on CK+ and we tested all 381 images of MMI for 7 classes. We achieved a recognition rate of 71.9 percent. However, the accuracy was only 64.11 percent when we performed training on MMI and testing on CK+. This is mainly due to fewer samples available for training and too many samples for testing. The recognition rates on cross databases are lower than those achieved when the same dataset was used for both training and testing. A reason behind this degradation is the diversity or the variations in terms of poses, control conditions and facial shapes among two databases and occlusions in MMI faces. However, the proposed method still performs better than the methods in [49] (51.1 percent) and [53] (65.47 percent).

4.4 Comparison with Recent Methods

We compared the performance of the proposed CCFS for 7-class recognition with recent methods for FER [17], [34], [49], [51], [52], [53], [54], [55] in Table 8. The lack of information in evaluation protocols and experimental settings makes the comparison task difficult, however, we present the comparison with methods that followed similar experimental settings. More specifically, average recognition rates of 91.79 percent for the JAFFE database and 94.09 percent for the CK+ database using 10-fold cross validation have been reported recently in [51], which is more closely related to our work. The author in [51] identified salient patches, extracted LBP features and used PCA+LDA before SVM for 6-class recognition. The proposed methodology based on feature search performed better than this patch based method. The patch based method heavily relies on the optimal size and location of each patch, which is hard to determine. Different facial muscles may have different sizes and the patch sizes may also vary in different areas of the face. Similarly, the method in [52] makes use of Adaboost to search for the discriminant Gabor features for six emotions from face regions. However, the spatial locations learned from these patches were not consistent with other databases. Dimensionality reduction can also be performed for feature selection. In [34], an unsupervised feature selection method is developed by computing the linear graph embedding and then selecting the features based on L1-norm

TABLE 8
Comparison with Recent Methods in Fer

Ref.	Feature Extraction/Selection	Classes	Classifier	Recognition rate (%)		
				Jaffe	CK+	MMI
[49]	Boosted LBP	7	SVM	81.0	91.40	86.90
[52]	Patch based Gabor	6	SVM	92.93	94.48	-
[53]	Intra-class variation	6	SRC	94.70	90.47*	93.81
[34]	Gabor- Feature selection	7	SVM	80.0	67.71	69.17†
[54]	AURF	6	SVM	-	92.22	69.88†
[55]	AUDN	6	SVM	-	93.70	**75.85**†
[17]	LBP patches-MTSL	6	SVM	-	91.53	73.53†
[51]	LBP patches + PCA-LDA	6	SVM	91.79	94.09	-
Proposed	CCFS	7	SVM	**95.31**	**96.05**	**94.09** / 74.63†

Six basic expression+Contempt. †Total sequences.

regularized least square assumption. However, the performance of this method on Gabor features were noticeably degraded as the resulted descriptive features were insufficient for recognizing expressions. Our method in this case is efficient for local pattern learning that jointly exploits the co-cluster information from class samples and features, especially in high dimensional Gabor feature space.

In terms of salient regions, the proposed method also showed satisfactory performance when compared to deep learning based facial feature representation [54] and [55]. While using the fewer sequences on MMI similar to [49], the proposed method achieves higher recognition rate. By considering all frontal view image sequences, deep learning based method [55] showed the highest rate on 6 expressions. In our experiments, we also included the Neutral expression. When Neutral expression is taken into account, there may be degradation in performance in cases where the expressions are too mild to differentiate. However, this effect is not being considered by previous works, including [51]. It can be observed that the proposed CCFS consistently achieves the highest recognition rate on Jaffe and CK+ and comparable to [55] on MMI. The main reason for the superior performance of our method is that the features used for classification are selected from the entire set of original features. The selected features preserve the salient information that can discriminate the expressions among various subjects.

5 Conclusion

Gabor wavelets are useful for facial expression recognition, but the number of features is very high as compared to the number of available samples. Among these overwhelming number of features, only a small fraction may be needed. Feature selection in this regard improves the scalability (defying the curse of dimensionality) and reduces the identity bias when specific regions are located.

In this paper, a novel method for feature selection in facial expression recognition based on co-clustering is proposed. Our method is able to reveal the saliency of face regions by selecting a small number of dominant features from Gabor wavelets features.

Experiments have been conducted on widely used benchmark facial expression image databases. With the JAFFE and MMI databases, we achieved recognition rates of 95.31 and 94.09 percent respectively using SVM on reduced feature sets. With the CK+ database, we obtained the accuracy of 96.05 percent by using only 20 percent of the features. An interesting aspect of this work is that when a feature set containing as little as 0.9 percent of the entire feature data were used, we could still obtain an accuracy of 90.23 percent.

Several other feature selection approaches were also considered to detect prominent features in comparison with CCFS. The feature selection methods based on feature ranking analyze the features separately and perform selection one after another. This might be undesirable in multiclass recognition, where different features have differenet strength on discriminating classes. As comparison with other approaches, CCFS achieves better results, which indicates that the local structures in the feature matrix are more crucial for discriminant feature selection and is a better way to analyze class features jointly.

Making use of the discriminating ability of CCFS in terms of recognizing facial expressions in frontal view images, we can extend our work on data captured in the wild, in order to test and improve the robustness of proposed method on spontaneous, non-posed and multiview images. Performance on cross databases, variations in expressions within the same classes and face occlusions are the main challenges that need to be addressed. We can fuse geometric and appearance features along with feature selection to tackle these problems.

Acknowledgments

This work is supported by the Hong Kong Research Grants Council (Projects CityU 11214814 and C1007-15G). Sheheryar Khan is the corresponding author. A preliminary version of this work appears in [10] 16th IEEE International conference on Machine Learning and Cybernetics (ICMLC 2016).

References

[1] M. S. Bartlett, J. C. Hager, P. Ekman, and T. J. Sejnowski, "Measuring facial expressions by computer image analysis," *Psychophysiology*, vol. 36, pp. 253–263, 1999.

[2] P. Ekman and W. V. Friesen, *Emotion in the Human Face*. Englewood Cliffs, NJ, USA: Prentice-Hall, 1975.

[3] Y. Tian, T. Kanade, and J. Cohn, *Handbook of Face Recognition*, Berlin, Germany: Springer, 2005, pp. 247–275.

[4] M. Valstar, I. Patras, and M. Pantic, "Facial action unit detection using probabilistic actively learned support vector machines on tracked facial point data," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshop*, 2005, pp. 76–84.

[5] Z. Zhang, M. J. Lyons, M. Schuster, and S. Akamatsu, "Comparison between geometry-based and Gabor-wavelets-based facial expression recognition using multi-layer perceptron," in *Proc. IEEE Int. Conf. Automat. Face Gesture Recognit.*, 1998, pp. 454–459.

[6] G. Donato, M. Bartlett, J. Hager, P. Ekman, and T. Sejnowski, "Classifying facial actions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 21, no. 10, pp. 974–989, Oct. 1999.

[7] M. S. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, and J. Movellan, "Recognizing facial expression: Machine learning and application to spontaneous behavior," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2005, pp. 568–573.

[8] M. A. Amin and H. Yan, "An empirical study on the characteristics of Gabor representations for face recognition," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 23, no. 3, pp. 401–431, 2009.

[9] M. J. Lyons, J. Budynek, and S. Akamatsu, "Automatic classification of single facial images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 21, no. 12, pp. 1357–1362, Dec. 1999.

[10] S. Khan, L. Chen, X. Zhe, and H. Yan, "Feature selection based on coclustering for effective facial expression recognition," in *Proc. 16th. IEEE Int. Conf. Mach. Learn. Cybern.*, 2016, pp. 48–53.

[11] M. Suwa, N. Sugie, and K. Fujimora, "A preliminary note on pattern recognition of human emotional expression," in *Proc. Int. Joint Conf. Pattern Recognit.*, 1978, pp. 408–410.

[12] E. Sariyanidi, H. Gunes, and A. Cavallaro, "Automatic analysis of facial affect: A survey of registration, representation, and recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 6, pp. 1113–1133, Jun. 2015.

[13] H. Zhao, A. W. C. Liew, X. Xie, and H. Yan, "A new geometric biclustering algorithm based on the Hough transform for analysis of large-scale microarray data," *J. Theoretical Biol.*, vol. 251, no. 3, pp. 264–274, 2008.

[14] V. Vapnik, *Statistical Learning Theory*. New York, NY, USA: Wiley, 1998.

[15] H. Zhao, D. D. Wang, L. Chen, X. Liu, and H. Yan, "Identifying multi-dimensional co-clusters in tensors based on hyperplane detection in singular vector spaces," *PLOS One*, vol. 11, no. 9, 2016, Art. no. e0162293.

[16] P. Khorrami, T. Paine, and T. Huang, "Do deep neural networks learn facial action units when doing expression recognition?'' in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, 2015, pp. 19–27.

[17] L. Zhong, et al., "Learning multiscale active facial patches for expression analysis," *IEEE Trans. Cybern.*, vol. 45, no. 8, pp. 1499–1510, Aug. 2015.

[18] C. Shan, S. Gong, and P. W. McOwan, “A comprehensive empirical study on linear subspace methods for facial expression analysis,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshop*, 2006, pp. 153–153.

[19] Y. Zhang and Q. Ji, “Active and dynamic information fusion for facial expression understanding from image sequences,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 5, pp. 1–16, May 2005.

[20] M. Pantic and L. Rothkrantz, “Expert system for automatic analysis of facial expression,” *Image Vis. Comput.*, vol. 18, no. 11, pp. 881–905, 2000.

[21] L. Wolf and A. Shashua, “Feature selection for unsupervised and supervised inference: The emergence of sparsity in a weight-based approach,” *J. Mach. Learn. Res.*, vol. 6, pp. 1855–1887, Nov. 2005.

[22] C. W. Hsu and C. J. Lin, “A comparison of methods for multiclass support vector machines,” *IEEE Trans. Neural Netw.*, vol. 13, no. 2, pp. 415–425, Mar. 2002.

[23] M. Pantic and L. J. M. Rothkrantz, “Facial action recognition for facial expression analysis from static face images,” *IEEE Trans. Syst. Man Cybern.*, vol. 34, no. 3, pp. 1449–1461, Jun. 2004.

[24] M. Pantic and I. Patras, “Dynamics of facial expression: Recognition of facial actions and their temporal segments from face profile image sequences,” *IEEE Trans. Syst. Man Cybern.*, vol. 36, no. 2, pp. 433–449, Apr. 2006.

[25] W. H. Yang, D. Q. Dai, and H. Yan, “Finding correlated biclusters from gene expression data,” *IEEE Trans. Knowl. Data Eng.*, vol. 23, no. 4, pp. 568–584, Apr. 2011.

[26] M. Turk and A. P. Pentland, “Face recognition using eigenfaces,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 1991, pp. 586–591.

[27] M. S. Bartlett, J. R. Movellan, and T. J. Sejnowski, “Face recognition by independent component analysis,” *IEEE Trans. Neural Netw.*,vol. 13, no. 6, pp. 1450–1464, Nov. 2002.

[28] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, “Eigenfaces versus. fisherfaces: Recognition using class specific linear projection, *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 7, pp. 711–720, Jul. 1997.

[29] J. Weston and C. Watkins, “Multi-class support vector machines,” Cornell Univ., Ithaca, NY, USA, Tech. Rep. CSD-TR 98-04, 2004.

[30] P. Hall, J. S. Marron, and A. Neeman, “Geometric representation of high dimension, low sample size data,” *J. Royal Statistical Society: Series B (Statistical Methodology)*, vol. 67, no. 3, pp. 427–444, 2005.

[31] S. Busygin, O. A. Prokopyev, and P. M. Pardalos, “Biclutering in data mining,” *Comput. Operations Res.*, vol. 35, pp. 2964–2987, 2008.

[32] A. Shabalin, V. Weigman, C. Perou, and A. Nobel, “Finding large average submatrices in high dimensional data,” *Ann. Appl. Statist.*, vol 3, pp. 985–1012, 2009.

[33] B. B. Liu, C. W. Yu, D. Z. Wang, R. C. C. Cheung, and H. Yan, “Design exploration of geometric biclustering for microarray data analysis in data mining,” *IEEE Trans. Parallel Distrib. Syst.*, vol. 25, no. 10, pp. 2540–2550, Oct. 2014.

[34] L. Wang, K. Wang, and R. Li, “Unsupervised feature selection based on spectral regression from manifold learning for facial expression recognition,” *IET Comput. Vis.*, vol. 9, no. 5, pp. 655–662, 2015.

[35] I. Cohen, N. Sebe, A. Garg, L. Chen, and T. S. Huang, “Facial expression recognition from video sequences: Temporal and static modeling,” *Comput. Vis. Image Understanding*, vol. 91, pp. 160–187, 2003.

[36] A. A. Shabalin, et al., “Finding large average submatrices in high dimensional data,” *Ann. Applied Statistics*, vol. 3, pp. 985–1012, 2009.

[37] S. Madeira and A. Oliveira, “Biclustering algorithms for biological data analysis: A survey,” *IEEE Trans. Comput. Biol. Bioinf.*, vol. 1, no. 1, pp. 24–45, Jan.-Mar. 2004.

[38] S. Busygin, O. A. Prokopyev, and P. M. Pardalos, “Feature selection for consistent biclustering via fractional 0-1 programming,” *J. Combinatorial Optimization*, vol. 10, pp. 7–21, 2005.

[39] X. He, D. Cai, and P. Niyogi, “Laplacian score for feature selection,” in *Proc. Advances Neural Inf. Process. Syst.*, 2006, pp. 507–514.

[40] H. Zeng and Y.-M. Cheung, “Feature selection and kernel learning for local learning-based clustering,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 8, pp. 1532–1547, Aug. 2011.

[41] G. Roffo, S. Melzi, and M. Cristani, “Infinite feature selection,” in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 4202–4210.

[42] P. Viola and M. Jones, “Robust real-time object detection,” *Int. J. Comput. Vis.*, vol. 57, no. 2, pp. 137–154, 2001.

[43] T. Kanade, J. Cohn, and Y. Tian, “Comprehensive database for facial expression analysis,” in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit.*, 2000, pp. 46–53.

[44] T. S. Lee, “Image representation using 2D gabor wavelets,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vo. 18, no. 10, pp. 959–971, Oct. 1996.

[45] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba, “Coding facial expressions with gabor wavelets,” in *Proc. 3rd. Int. Conf. Face Gesture Recognit.*, 1998. pp. 200–205.

[46] D. Z. Wang and H. Yan, “A graph spectrum based geometric biclustering algorithm,” *J. Theoretical Biol.*, vol. 317, pp. 200–211, 2013.

[47] J. Movellan, “Tutorial on Gabor filters,” Tech. Rep. MPLab Tutorials, Univ. of California, San Diego, 2005.

[48] M. F. Valstar and M. Pantic, “Induced disgust, happiness and surprise: An addition to the MMI facial expression database,” in *Proc. Int. Conf. Lang. Resour. Eval.*, 2010, pp. 65–70.

[49] C. Shan, S. Gong, and P. W. McOwan, “Facial expression recognition based on local binary patterns: A comprehensive study,” *Image Vis. Comput.*, vol. 27, no. 6, pp. 803–816, 2009.

[50] M. H. Siddiqi, et al., “Human facial expression recognition using stepwise linear discriminant analysis and hidden conditional random fields,” *IEEE Trans. Image Process.*, vol. 24,.no. 4, pp. 1386–1398, Apr. 2015.

[51] S. L. Happy and A. Routray, “Automatic facial expression recognition using features of salient facial patches,” *IEEE Trans. Affect. Comput.*, vol. 6, no. 1, pp. 1–12, Jan.-Mar. 2015.

[52] L. Zhang and D. Tjondronegoro, “Facial expression recognition using facial movement features,” *IEEE Trans. Affect. Comput.*, vol. 2, no. 4, pp. 219–229, Oct.-Dec. 2011.

[53] S. H. Lee, K. N. Plataniotis, and Y. M. Ro, “Intra-class variation reduction using training expression images for sparse representation based facial expression recognition,” *IEEE Trans. Affect. Comput.*, vol. 5, no. 3, pp. 340–351, Jul.-Sep. 2014.

[54] M. Liu, et al., “Au-aware deep networks for facial expression recognition,” in *Proc. 10th IEEE Int. Conf. Workshops Autom. Face Gesture Recognit.*, 2013, pp. 1–6.

[55] M. Liu, et al., “Au-inspired deep networks for facial expression feature learning,” *Neurocomputing*, vol. 159, pp. 126–136, 2015.

Sheheryar Khan received the BS degree from UET Peshawar, Pakistan with Honours, in 2008 and the MSc degree from Lancaster University, United Kingdom with distinction in 2010. Before joining City University of Hong Kong as a PhD candidate in 2015, he was as a lecturer in COMSATS Institute of Information Technology, Pakistan. His research interests include image processing, pattern recognition and facial expression analysis. He is a student member of the IEEE.

Lijiang Chen received the BS and PhD degrees from School of Electronic and Information Engineering, Beihang University, Beijing, China, in 2007 and 2012 respectively. He is currently a lecturer in the School of Electronic and Information Engineering, Beihang University, Beijing, China. His research interests include speech signal processing, pattern recognition and speech emotion analysis. He is a member of the IEEE.

Hong Yan received the PhD degree from Yale University. He was professor of imaging science with the University of Sydney and currently is professor of computer engineering with City University of Hong Kong. His research interests include image processing, pattern recognition and bioinformatics. He has authored or co-authored more than 300 journal and conference papers in these areas. He was elected an IAPR fellow for contributions to document image analysis and an IEEE fellow for contributions to image recognition techniques and applications.