Multimed Tools Appl (2016) 75:935–959
DOI 10.1007/s11042-014-2333-3

Human facial expression recognition using curvelet feature extraction and normalized mutual information feature selection

Muhammad Hameed Siddiqi · Rahman Ali · Muhammad Idris · Adil Mehmood Khan · Eun Soo Kim · Min Cheol Whang · Sungyoung Lee

Received: 2 May 2014 / Revised: 26 August 2014 / Accepted: 20 October 2014 /
Published online: 5 November 2014

Abstract To recognize expressions accurately, facial expression systems require robust feature extraction and feature selection methods. In this paper, a normalized mutual information based feature selection technique is proposed for FER systems. The technique is derived from an existing method, that is, the max-relevance and min-redundancy (mRMR) method. We, however, propose to normalize the mutual information used in this method so that the domination of the relevance or of the redundancy can be eliminated. For feature extraction, curvelet transform is used. After the feature extraction and selection the feature space is reduced by employing linear discriminant analysis (LDA). Finally, hidden Markov model (HMM) is used to recognize the expressions. The proposed FER system (CNF-FER) is validated using four publicly available standard datasets. For each dataset, 10-fold cross validation scheme is utilized. CNF-FER outperformed the existing well-known statistical and state-of-the-art methods by achieving a weighted average recognition rate of 99 % across all the datasets.

Keywords Facial Expressions · Curvelet Transform · Mutual Information · Minimal Redundancy · Maximal Relevance

M. H. Siddiqi · R. Ali · M. Idris · S. Y. Lee (✉)
Department of Computer Engineering, Kyung Hee University,
Suwon 446–701, Republic of Korea
e-mail: sylee@oslab.khu.ac.kr

A. M. Khan
Division of Information and Computer Engineering, Ajou University,
Suwon 443–749, Republic of Korea

E. S. Kim
Department of Electronic Engineering, Kwangwoon University,
Seoul 139–701, Republic of Korea

M. C. Whang
Division of Digital Media Engineering, Sang-Myung University,
Suwon 110–809, Republic of Korea

1 Introduction

Communication through facial expressions plays a significant role in social interactions. Over the past two decades, human facial expression recognition (FER) emerged as an important research area. FER systems are very important to applications such as human emotion analysis [11], psychology and cognitive sciences [35], and access control and surveillance [6]. In most of these applications, FER systems should be adequately intelligent such that they can easily understand and analyze the goals and behaviors of humans.

There are two categories of FER systems: posed FER systems [39–41], and spontaneous FER systems [1, 17, 31]. Posed FER systems are used for recognizing artificial expressions, that is, the expressions produced by people when they are asked to do so [6]. On the other hand, spontaneous expressions are those that are produced by people spontaneously; these are observed on daily basis, such as during conversation or while watching movies [6]. This work falls under the category of posed FER systems.

Usually, an FER system consists of four basic modules: preprocessing, feature extraction, feature selection, and recognition. The preprocessing module is used to improve the image quality by diminishing the illumination noise and by eliminating the unnecessary details from the background. Feature extraction module deals with getting the distinguishable features for each expression and quantizing them as discrete symbols. Feature selection module is used for selecting a subset of relevant features from a large number of features extracted from the input data. Finally, in the recognition module, a classifier is first trained using the training data and then used to generate labels for the expressions in the incoming video data.

Many well-known methods, such as Histogram Equalization (HE) [33], Weighted Vector Directional Filters (WVDF) [18] or Wiener filter [21] have been employed for preprocessing. Similarly, for feature extraction, a number of techniques have been developed. Among these techniques, curvelet transform is the most robust and accurate technique. The accuracy and robustness of the curvelet transform has been proven by [42]. Its efficiency in coding the image edge information is high [8].Therefore, curvelet transform [42] was chosen for feature extraction in this work. For further study, please refer to [42].

Likewise, for the recognition module, many well-known classifiers have been studied for accurate expressions classification. For instance, artificial neural networks (ANNs) were employed by [12], support vector machines SVMs by [22, 38], Gaussian mixture models GMMs by [37] and hidden Markov models HMMs by [29]. Among these classifiers, HMM is the most frequently employed and commonly tested technique for sequential data [43].

In pattern recognition, identification of the most discriminative features is an important step [9], since it is common to have a large number of features, including relevant as well as irrelevant features, at the beginning of the pattern recognition process [13, 16]. Feeding a large set of features into a recognition model not only increases the computation burden but also causes the problem commonly known as the curse of dimensionality. Therefore, selecting only the relevant features helps in speeding up the learning process and alleviates the affect of the curse of dimensionality. Furthermore, feature selection facilitates in data visualization and understanding [19]. In regard to feature selection for FER, a number of techniques have been investigated. Among these, the most commonly used method is the mutual information based feature selection. However, there are still some limitations in this method [2, 10, 23, 30]. For instance, given a dataset with N features $X_1, X_2, ..., X_N$, and a set of $i-1$ selected index ($S_{i-1} = \{s_1, s_2, ..., s_{i-1}\}$), the next feature X_{s_i} is selected

so that the redundancy $\left(RC\left(X_{s_i}\right) = \sum_{s \in s_{i-1}} I\left(X_s; X_{s_i}\right)\right)$ is minimized and the relevance $\left(RL\left(X_{s_i}\right) = I\left(C; X_{s_i}\right)\right)$ is maximized. However, because the two problems may not have a common solution; therefore, we would like to find a scalar factor (denoted by β) so that a feature X_{s_i} maximizing $RL\left(X_{s_i}\right) - \beta \times RD\left(X_{s_i}\right)$ will be a possible solution for the minimization as well as the maximization. The existing solutions are summarized as given below:

- MIFS and and MIFS-U: β is manually selected by the experiments,
- mRMR: $\beta = \frac{1}{S_{i-1}}$.
- NMIFS: $\beta\left(X_s; X_{s_i}\right) = \frac{1}{S_{i-1}} \times \frac{1}{\min\left(H(X_s), H\left(X_{s_i}\right)\right)}$.

where MIFS stands for mutual information feature selection, MIFS-U stands for mutual information feature selection-unsupervised, mRMR stands for max-relevance and min-redundancy, and NMIFS stands for normalized mutual information feature selection.

Therefore, the objective of this paper is to propose a robust feature selection method in which we utilize the information measurement in order to estimate the potential of the features. On the topic of searching algorithms, since an exhaustive search over a large feature space is impractical, greedy forward selection and backward elimination are often used [2, 23, 30]. Here, we exploit greedy forward selection, wherein each feature is appended to the feature set based on its quality. Moreover, in this research, a detailed study on curvelet transform in combination with the proposed feature selection method is performed in order to extract and select the most prominent features. The dimension of the feature space is further reduced by employing a well-known statistical method called linear discriminant analysis LDA, and finally, HMM is used to label the expressions. System vlidation is perfmorned on four publicly available standard datasets of facial expressions, i.e., Japanese Female Facial Expressions (JAFFE) dataset [26], Yale B face dataset [14], Cohen-Kanade dataset [20], and Natural Visible and Infrared Facial Expression (USTC-NVIE) dataset[44].

We already discussed some related work about this field. The rest of the paper is organized as follows. Section 2 provides an overview of the proposed FER system (CNF-FER) with the integration of LDA and HMM. The experimental setup for the CNF-FER system is described in Section 3. Section 4 presents the experimental results along with discussion on each experiment, comparing our results with other well-known and state-of-the-art methods. Finally, Section 5 provides the conclusion of the paper with some future directions.

2 Methodology

2.1 Feature extraction

As mentioned above, curvelet transform [42] is used for feature extraction. This method has the capability to extract the most prominent features by keeping the line, curve, and edge information from each expression frame. It is a very light method for scenarios where objects edge information is illustrated. It can also be used for image reconstruction in severely ill-posed problems.

Curvelet transform can be implemented in two ways: firstly, using Unequally Spaced Fast Fourier Transform (USFFT); and secondly, through wrapping. We utilized curvelets through wrapping because it is faster than the USFFT method. While reconstructing the edge

details in an image, this method is capable of employing a small number of coefficients, comparatively. Coefficient matrices of angle and scale are given as:

$$C\left(j,l,k\right)=\left\langle f,\varphi_{j,l,k}\right\rangle \tag{1}$$

where j represents the scale, l indicates the angle, k shows the parameter position, and the inner products project f onto the $\varphi_{j,l,k}$, which represents the basic function of the curvelet transform. The angle and scale are indicated in Fig. 1.

As described before, we employed curvelet via wrapping; therefore, the following steps are used for it [7].

- In first step, fast Fourier transform is employed to get the Fourier samples. $\hat{f}\left[n_1,n_2\right], -\frac{n}{2}\le n_1,n_2\le\frac{n}{2}$ (see Fig. 1).
- In second step, the interpolation (re-sample) $\hat{f}\left[n_1,n_2\right]$ takes place for each pair of scale j and angle l in order to attain $\hat{f}\left[n_1,n_2-n_1\tan\theta_l\right]$.
- In the third step, multiplication of interpolation function $\hat{f}$ with a discrete localizing window function is performed $\tilde{U}_j\left[n_1,n_2\right]$. $\hat{f}_{j,1}\left[n_1,n_2\right]=\hat{f}\left[n_1,n_2-n_1\tan\theta_l\right]\tilde{U}_j\left[n_1,n_2\right]$
- In last step, on each $\hat{f}_{j,1}$, the inverse fast Fourier transform is performed in order to attain the associated curvelets $C\left(j,l,k\right)$.

where the range of n_1 and n_2 is between $0\le n_1\le L_{1,j}$ and $0\le n_2\le L_{2,j}$, while the range of θ is between $-\frac{\pi}{4}$ and $\frac{\pi}{4}$. For more detail on curvelet transform, please refer to [7, 42].

2.2 Normalized mutual information-based feature selection (NMIFS)

As mentioned before, for feature selection module, a robust normalized mutual information feature selection technique is used. This method is derived from the max-relevance and min-redundancy (mRMR) approach. We, however, propose to normalize the mutual information used in the method so that the domination of the relevance or of the redundancy can be eliminated. In our method only the upper bound of the mutual information of random variables is considered. Since, any continuous variable can be quantized into a discrete form, it is assumed that two discrete random variables X and Y are given along with

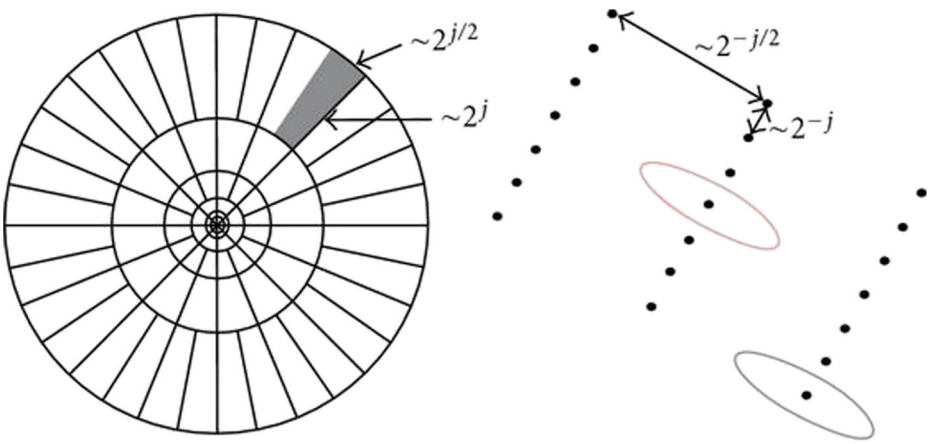

Fig. 1 The illustration of curvelet transform in frequency domain (*left*) and in spatial domain (*right*) [7]

their marginal and joint distributions. Hence, the joint mutual information of X and Y is computed as:

$$I(X;Y) \leq \min(H(X), H(Y)) \tag{2}$$

where I is the joint mutual information of two random variables X and Y, while, H is the entropy function, which can be defined by employing Jensen's inequality that is given as

$$H(X) \leq \log_2\left(\sum_{x\in\Omega_X} p(x)\frac{1}{p(x)}\right) \tag{3}$$

$$H(X) \leq \log_2(|\Omega_X|) \tag{4}$$

It is clear from (2) and (4) that

$$I(X;Y) \leq \min\left(\log_2(|\Omega_X|), \log_2(|\Omega_Y|)\right) \tag{5}$$

In this work, every feature is quantized by employing the same number of levels (N) that has been decided in order to achieve the expected quantization error. Algorithm 1 illustrates the quantization algorithm.

Algorithm 1 Quantization algorithm.

```
Input  : M−Total number of features
         X(1..M)−Training data
         ξ−The quantization error
Output: N−Number of quantization levels
         Y(1..M)− Quantized data

begin
    N = 2
    while 1 do
        ManError = −1e + 16 for m = 1 to M do
            Upper = max(X(m)) Lower = min(X(m))
            Step = (Upper − Lower)/N Partition = [Lower : Step : Upper]
            CodeBook = [Lower − Step, Lower : Step : Upper]
            [Y(m), QError] = Quantiz(X(m)), Partition, CodeBook if
            QError > MaxError then
                MaxError = QError
            end
        end
        if MaxError < ξ then
            Break
        end
        N = N + 1
    end
end
```

It is clear that the number of quantization levels progressively increases until the quantization error becomes smaller than a predefined small constant (ξ) that is the expected quantization error. We used $\xi = 0.05$ for our experiments because smaller values than this did not improve accuracy but increased the computational cost. It is clear from the algorithm that $|\Omega_X| = N$ for every feature X, so,

$$I(X;Y) \leq \log_2(N) \tag{6}$$

$\log_2(N)$ is the upper bound of the mutual information $I(X;Y)$ and does not depend either on X or Y; therefore, we call $\log_2(N)$ as a feature-independent upper bound.

To eliminate the problem of unequal normalizing weights, we propose to use the feature-independent upper bound in (6) to normalize the mutual information instead of

employing (2) as in [10]. Therefore, our normalized feature-feature mutual information is calculated by

$$NI\left(X;Y\right)=\frac{I\left(X;Y\right)}{\log_2\left(N\right)} \tag{7}$$

It is clear that the range for the normalized feature-to-feature mutual information is always within [0, 1]. Therefore, the class-feature mutual information is divided by $\log_2|\Omega_C|$ in order to achieve a balance between the relevance and the redundancy, which is now defined as

$$NI\left(C;X\right)=\frac{I\left(C;X\right)}{\log_2|\Omega_C|} \tag{8}$$

Combining (6) and (7), the potential of a feature is measured as

$$f^1\left(X_i\right)=NI\left(C;X_i\right)-\frac{1}{|S_{i-1}|}\sum_{X_s\in S_{i-1}}NI\left(X_s;X_i\right) \tag{9}$$

where, S is a feature set, i.e., $S=X_1, X_2, ..., X_i$. In addition, to validate the effect of the imbalance between the relevance and the redundancy that we pointed out above, the normalized class-feature mutual information is combined with the same feature-feature mutual information as in [10]. In this way, the goodness of the feature is measured by

$$f^1\left(X_i\right)=NI\left(C;X_i\right)-\frac{1}{|S_{i-1}|}\sum_{X_s\in S_{i-1}}\frac{I\left(X_s;X_i\right)}{\min\left(H\left(X_s\right),H\left(X_i\right)\right)} \tag{10}$$

The following pseudo code in Algorithm 2 represents the selection process using greedy forward searching strategy.

Algorithm 2 Mutual information-based feature selection using greedy forward searching.

Input : $M-$Total number of features
$N-$Total number of data samples
$K-$Number of features to be selected
$X_{ij}-$Feature values, where i=1,2,...,M and j=1,2,...,N
C_j-Class labels of the data samples, where j=1,2,...,N
$a-$Index of the selected measurement

Output: S_k-The selected feature index, where k=1,2,...,K

```
begin
    S = φ
    for m = 1 to M do
        μ_m = Mean value of X_m
        σ = Standard deviation of X_m
        X_m = X_m − μ_m
        X_m = X_m/σ_m
    end
    X̄ = Quantiz(X)
    for k = 1 to K do
        for i = 1 to M do
            Computef^a(X̄_i)
        end
        s = argmax_{i∉S}(f^a(X̄_i))
        S = S ∪ s
    end
end
```

2.3 Dimension reduction

Some well-known methods used for dimension reduction of a feature space include kernel discriminant analysis (KDA) [28], generalized discriminant analysis (GDA) [3], and LDA [27]. Among these, LDA has been widely used in FER domain.

2.3.1 Linear discriminant analysis

Linear discriminant analysis maximizes the ratio of between-class variance to within-class variance in any particular data set, thereby guaranteeing maximal separability. LDA produces an optimal linear discriminant function that maps the input into the classification space on which the class identification of the samples is decided. LDA easily handles the case in which the within-class frequencies are unequal. The within S_W and between S_B class comparison is done by using the following equations.

$$S_B = \sum_{i=1}^{c} V_i \left(\overline{m}_i - \overline{\overline{m}}\right)\left(\overline{m}_i - \overline{\overline{m}}\right)^T \tag{11}$$

$$S_W = \sum_{i=1}^{c} \sum_{m_k \in C_i} \left(m_k - \overline{m}_i\right)\left(m_k - \overline{m}_i\right)^T \tag{12}$$

where V_i is the number of vectors in the i_{th} class C_i, and c is the number of classes, and in our case, c represents the number of facial expressions. Also, $\overline{\overline{m}}$ represents the mean of all the vectors, $\overline{m}$ is the mean of the class C_i, and m_k is the vector of a specific class. The optimal discrimination projection matrix D_{opt} is chosen from the maximization of the ratio of determinant of the between and within-class scatter matrices as

$$D_{opt} = \arg\max_{D} \frac{\left|D^T S_B D\right|}{\left|D^T S_W D\right|} = [d_1, d_2, \ldots, d_t]^T \tag{13}$$

where D_{opt} is the set of discriminate vectors of S_W and S_W corresponding to the $c-1$ largest generalized eigenvalues λ. The size of D_{opt} is $t \times r$, where $t \leq r$, and r is the number of elements in a vector. Then,

$$S_B d_i = \lambda_i S_W d_i, i = 1, 2, ..., c-1 \tag{14}$$

where the rank of S_B is $c-1$ or less, and hence, the upper bound value of t is $c-1$. Thus, LDA maximizes the total scattering of the data while minimizing the within scattering of the classes. For more details on LDA, please refer to [5].

2.4 Expressions labeling using hidden Markov model

As described before, HMM is the most commonly used method for sequential data (facial expressions) classification, which provides a statistical model λ for a set of observation sequences. These observations are called frames in FER domain. A typical HMM has a sequence of observations of length T (i.e., $T = O_1, Q_2, ..., O_T$), a sequence of states S (i.e., $S = S_1, S_2, ..., S_N$, where N is the number of states in the model), and the time t for each state is denoted by Q (such that $Q = q_1, q_2, ..., q_N$). Each time, when a state

j is entered, an observation is generated according to the multivariate Gaussian distribution $b_j(O_t)$ with the mean value μ_j and covariance matrix V_j correlated with that state. There is also transition probabilities correlated with them such that the probability a_{ij} is the resultant transition probability from state i to state j. The initial model probability for the state j is Π_j. An HMM can be defined by this set of parameters, such as $\lambda = A, B, \Pi$, where A indicates the probability of the state transition (such that $A = a_{ij}$, $a_{ij} = Prob(q_{t+1} = S_j | q_t = S_i)$, $1 \leq i, j \leq N$), where B represents the probability of observations (such that $B = b_j(O_t)$, $b_j = Prob(O_t | q_t = S_j) 1 \leq j \leq N$), and the initial state probability is indicated by Π (such that $\Pi = \Pi_j$, $\Pi_j = Prob(q_1 = S_1)$). All the equations are based on the work by [32] and make use of the initial state probability distribution.

In the training step, for a given model λ, the multiplication of each transition probability by each output probability at each step t provides the joint likelihood of a state sequence Q and the corresponding observation O. This likelihood $P(O|\lambda)$ can be evaluated by summing over all possible state sequences:

$$P(O|\lambda) = \sum_{Q} P(O, Q|\lambda) \tag{15}$$

A simple procedure for finding the parameters λ that maximize the above equation in HMM, introduced in [4], depends on forward and backward algorithms $\alpha_t(j) = P(O_1 ... O_t, q_t = j|\lambda)$ and $\beta_t(j) = P(O(t+1) ... O_T | q_t = j, \lambda)$, respectively, such that these variables can be initiated inductively by the following processes:

$$\alpha_1(j) = \pi_j b_j(O_1), 1 \leq j \leq N \tag{16}$$

$$\beta_T(j) = 1, 1 \leq j \leq N \tag{17}$$

During testing, the appropriate HMM can then be determined by mean of likelihood estimation for the sequence observations O calculated based on the trained λ as

$$P(O|\lambda) = \sum_{i=1}^{N} \alpha_T(i) \tag{18}$$

The maximum likelihood for the observations provided by the trained HMM indicates the recognized label. The following formula has been utilized to model HMM (λ).

$$\lambda = (O, Q, \pi) \tag{19}$$

where O is the sequence of observations (i.e., $O_1, Q_2, ..., O_T$) and each state is denoted by Q (such as $Q = q_1, q_2, ..., q_N$), where N is the number of states in the model, and π is the initial state probabilities. The parameters that are used to model HMM (λ) for all experiments were 64, 4, and 4 respectively. These values have been selected by performing multiple experiments. For more details on HMM, please refer to [36].

3 Experimental setup

The CNF-FER is tested and validated on four publicly available standard datasets, namely JAFFE, Yale B, Cohn-Kanade, and USTC-NVIE datasets. Six basic universal expressions, that is, happiness, anger, sadness, disgust, surprise, and fear are used from these datasets. From each dataset, we have selected only those expression frames which display the frontal view of the face, and each expression is composed of several sequences of expression frames. In this work, 10-fold cross-validation scheme was applied, i.e., out of 10 subjects, data from a single subject was reserved as the validation data for testing the CNF-FER, whereas the data for the remaining 9 subjects were used as the training data. This process was repeated 10 times. There were some expressions in the datasets that have different lighting conditions; therefore, histogram equalization was used in order to diminish the lighting effects. The detail on each dataset is as follows:

- *JAFFE Dataset*:
  The expressions in this dataset were posed by 10 different subjects (Japanese female). Each image has been rated on 6 expression adjectives by 60 Japanese subjects. Most of the expression frames were taken from the frontal view of the camera with tied hair in order to expose all the sensitive regions of the face. In the whole dataset, there were total 213 facial frames, which consists of seven expressions including neutral. Therefore, we selected only 195 expression frames for six facial expressions performed by ten different Japanese female as subjects. The original size of each facial frame was 256×256 pixel.
- *Yale B Face Dataset*:
  There were total 5760 facial frames taken in single light source performed by 10 distinct subjects, each seen under 576 viewing conditions (9 poses $\times$ 64 illumination conditions). For every subject, while performing a particular pose, the ambient illumination was also captured.
- *Cohn-Kanade Dataset*:
  In this facial expressions dataset, 100 subjects (university students) performed basic six expressions. The age range of the subjects were from 18 to 30 years and most of them were female. We employed those expression frames for which the camera was fixed in front of the subjects. In order to utilize the six expressions from this dataset, we employed a total 450 image sequences from 100 subjects, and each of them was considered as one of the six expressions. The original size of each facial frame was 640×480 or 640×490 pixel with 8-bit precision for grayscale values. For recognition purpose, twelve expression frames were taken from each expression sequence, which results in total of 5400 expression images.
- *USTC-NVIE Dataset*:
  In this dataset, an infrared thermal and a visible camera was used in order to collect both spontaneous and posed expressions, but we utilized only posed-based expressions. There were 108 subjects (university students), and their age range was from 17 to 31 years.Some of them worn glasses, whereas others were free of glasses. They were asked to perform a series of expressions with illumination from three different directions. The size of each facial frame was 640×480 or 704×490 pixels. In total, 1027 expression frames were utilized from this dataset.

For a thorough validation, we performed four different sets of experiments in this study:

- In the first experiment, we analyzed the performance of the previous different statistical approaches with different combinations, using all datasets. Similarly, the performance of wavelet transform was also analysed in this experiments on all datasets.
- In the second experiment, performance of the CNF-FER was analyzed.
- While, in the third experiment, effectiveness of the proposed feature selection method was analyzed.
- Finally, In the last experiment, the weighted average recognition rate and time complexity of the CNF-FER were compared with some state-of-the-art methods.

4 Results and discussion

4.1 First experiment

In this experiment, four sub-experiments were performed based on the combination of different well-known techniques using all the datasets. Each experiment along with its results are described below:

- **Recognition Rates of PCA and LDA with HMM**:
  In this experiment, PCA was used along with LDA and HMM. PCA is an unsupervised technique, i.e., it does not require any prior information about the classes. For any

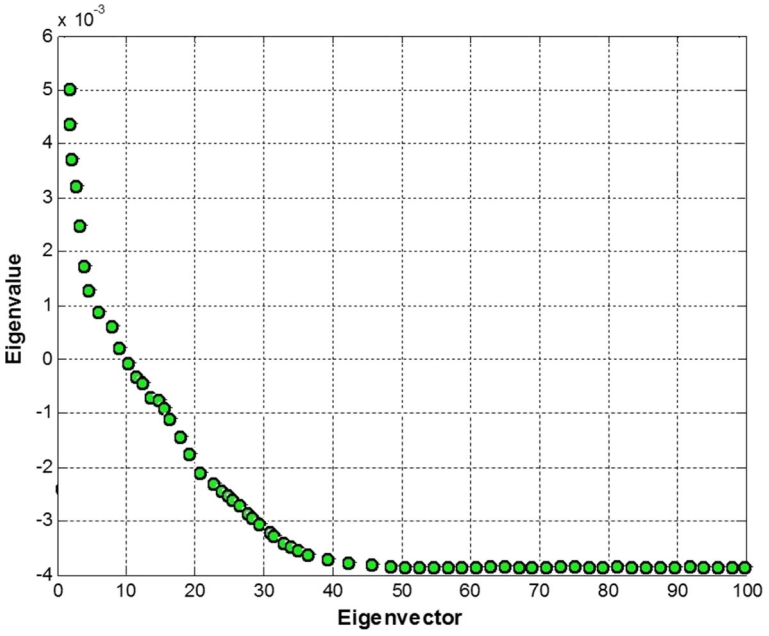

Fig. 2 Top 100 eigen values along with their eigenvectors using Cohn-Kanade dataset of facial expressions

defined level of compression, PCA is an optimal dimension-reduction scheme that minimizes the mean squared error between the original images and their reconstructions. Analysis showed that, when the number of PC features was increased, the recognition rate rose to a certain value and then remained saturated as shown in Fig. 2. Therefore, for optimal results, the first 50 eigenvectors with their corresponding eigenvalues were employed for each dataset. Once the features were extracted, LDA was applied and the LDA features were then fed to an HMM for recognition. The idea was to maximize the total scattering of the data while minimizing the variance within classes before recognition. The recognition rates for this experiment on the four datasets are shown in Fig. 3.

It is clear from Fig. 3 that PCA and LDA with HMM did not achieve high recognition rate. The reason for this could be that PCA only focuses on global features and does not capture the local features.

– **Recognition Rates of ICA and LDA with HMM**:
In this experiment, we used ICA for feature extraction (to capture the local features) with LDA in order to examine any improvement in the feature space. The results for this experiment are summarized in Fig. 4.

It can be seen from Fig. 4 that using the global or local features separately with LDA does not guarantee a better recognition rate. Because ICA is slow to train when the dimension of the data is bulky. Moreover, ICA is very weak in managing the inputs, e.g., if a hug amount of expressions frames are exploited as input, ICA does not have the capability to recognize it, due to which some time ICA cannot retrieve the desire features. Therefore, both, PCA and ICA with LDA can get most informative features.

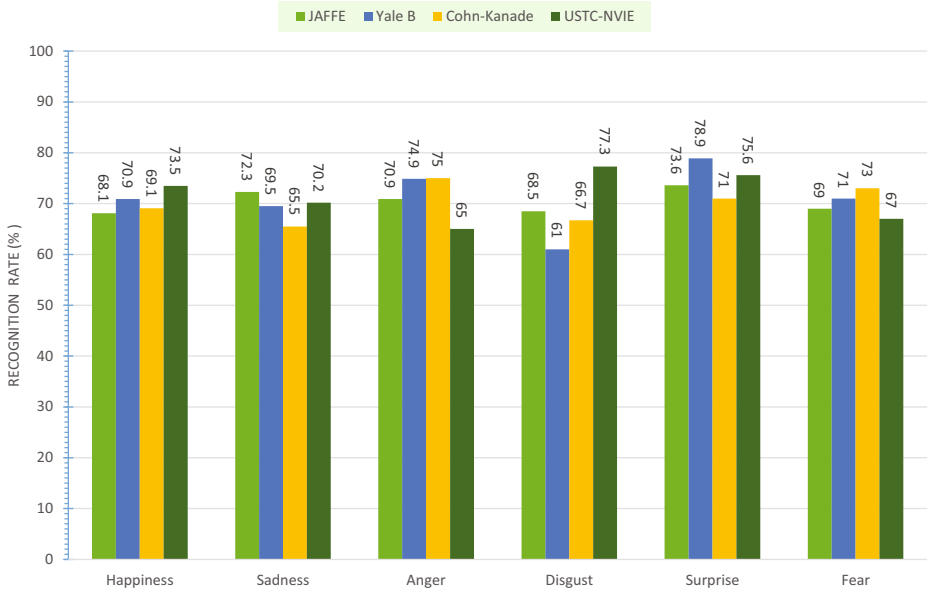

Fig. 3 Recognition rate of PCA and LDA with HMM using four datasets of facial expressions

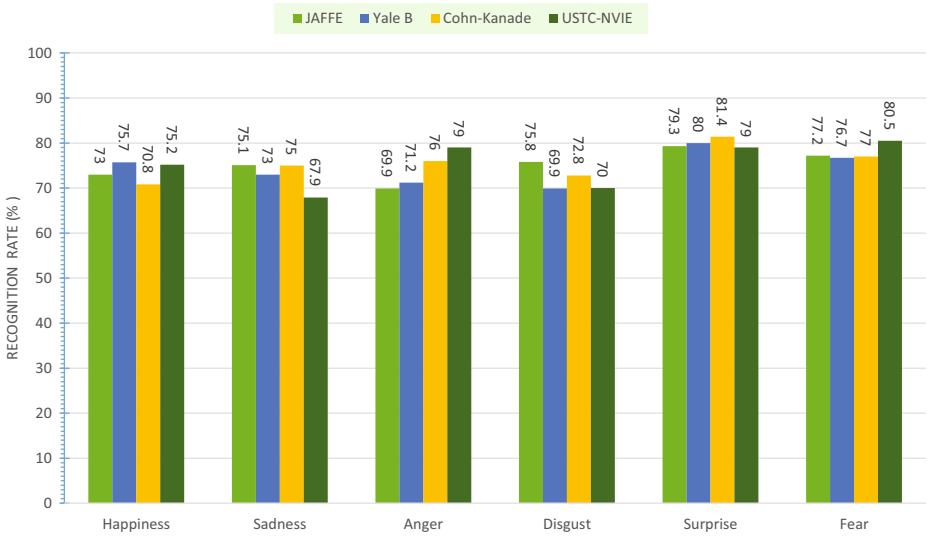

Fig. 4 Recognition rate of ICA and LDA with HMM using four datasets of facial expressions

– **Recognition Rates of PCA+ICA+LDA and HMM**:
In this experiment, we utilized both PCA and ICA to extract the global and local features with LDA and HMM. The results for this experiment are summarized in Fig. 5.

It can be seen from Fig. 5 that all the existing statistical methods with different combinations did not achieve better recognition rate due to their own limitations.

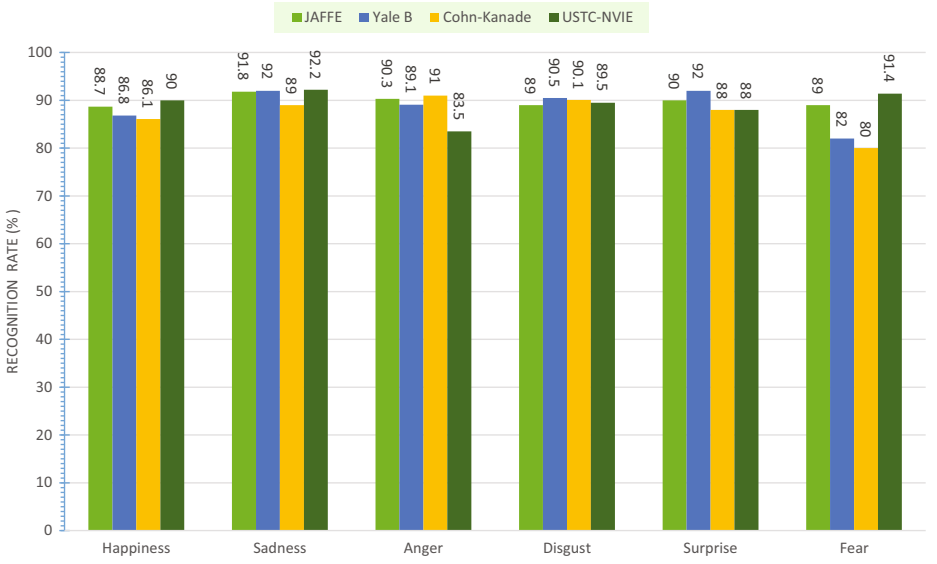

Fig. 5 Recognition rate of PCA+ICA and LDA with HMM using four datasets of facial expressions

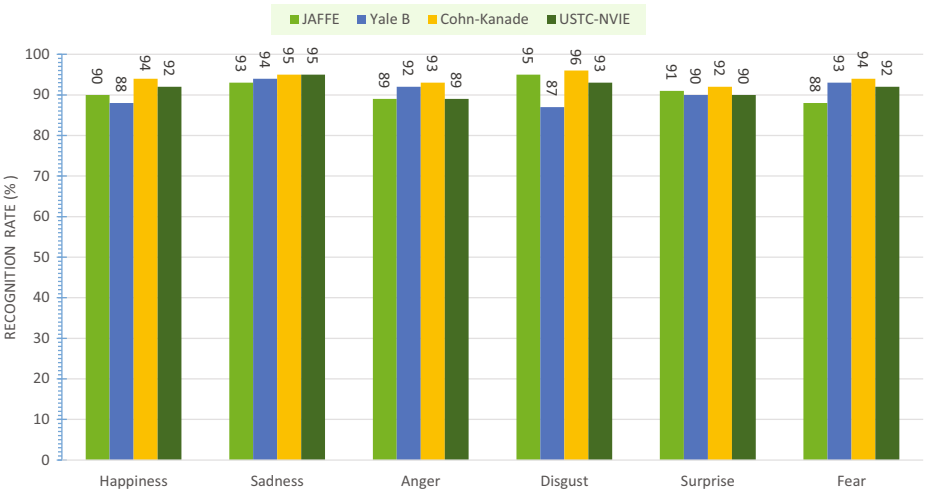

Fig. 6 Recognition rate of wavelet transform with HMM using four datasets of facial expressions

– **Recognition Rates of wavelet transform with HMM**:
Finally, we analyzed the accuracy of the wavelet transform as a feature extraction technique with LDA and HMM for expression recognition. The corresponding experimental results are indicated in Fig. 6.

It is clear from Fig. 6 that the wavelet transform does not achieve high recognition rate. Therefore, we proposed the CNF-FER in order to attain a better performance against the existing methods on all datasets.

4.2 Second experiment

In this experiment, two sub-experiments were performed to validate the CNF-FER using the four datasets. The same validation scheme was applied as mentioned in Section 3, and the results are described below.

– **Recognition Rates of CNF-FER**:
The CNF-FER was evaluated for each dataset separately under the exact settings as mentioned in Section 3. The 3D feature plots of the CNF-FER, for the six expressions, after applying the LDA on four datasets, are shown in Figs. 7, 8, 9, and 10, and the detailed results are provided in Table 1.

It is clear from Table 1 that the CNF-FER consistently achieved a high recognition rate when applied on these datasets separately, i.e., 99.00 % on JAFFE, 99.17 % on Yale B face dataset, 99.17 % on Cohn-Kanade, and 99.33 % on USTC-NVIE dataset. This means that, unlike existing statistical methods, such as PCA, ICA, LDA, and wavelet transform, the CNF-FER is more robust, i.e., it provided high recognition rate not just for one but all four datasets. This is due to the proposed feature selection method that utilizes the information measurement in order to estimate the potential of the features. Furthermore, greedy forward selection was used, wherein each feature is appended to the feature set based on its quality, which confirms that the proposed feature selection method is more robust than others with respect to classification accuracy.

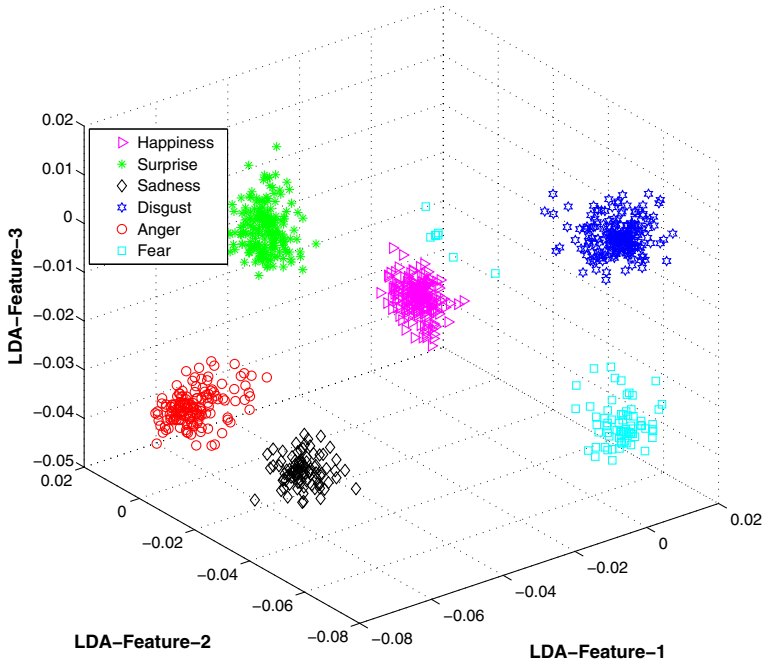

Fig. 7 3D feature plots of the CNF-FER for recognizing the expressions (on JAFFE dataset). It can be seen that the CNF-FER clearly classified the expressions classes

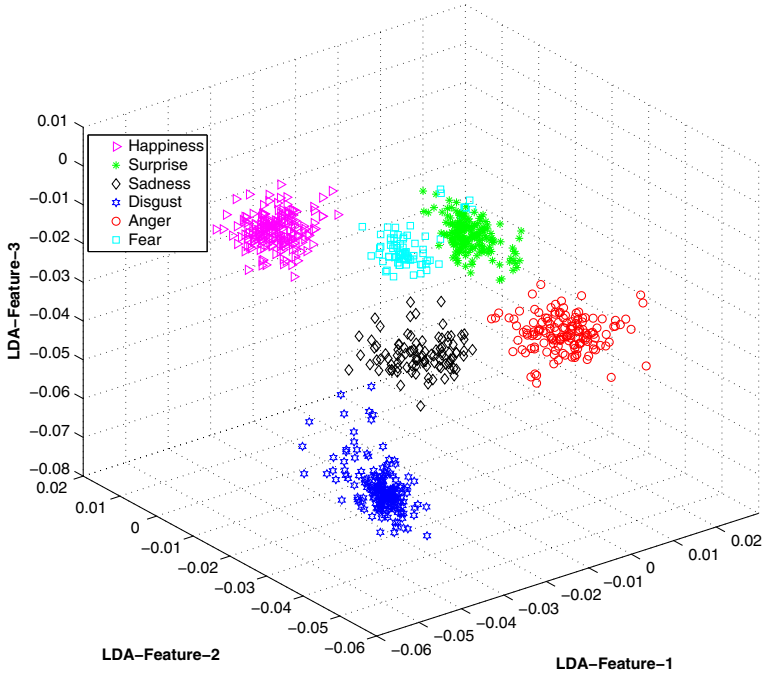

Fig. 8 3D feature plots of the CNF-FER for recognizing the expressions (on Yale B face dataset). It can be seen that the CNF-FER clearly classified the expressions classes

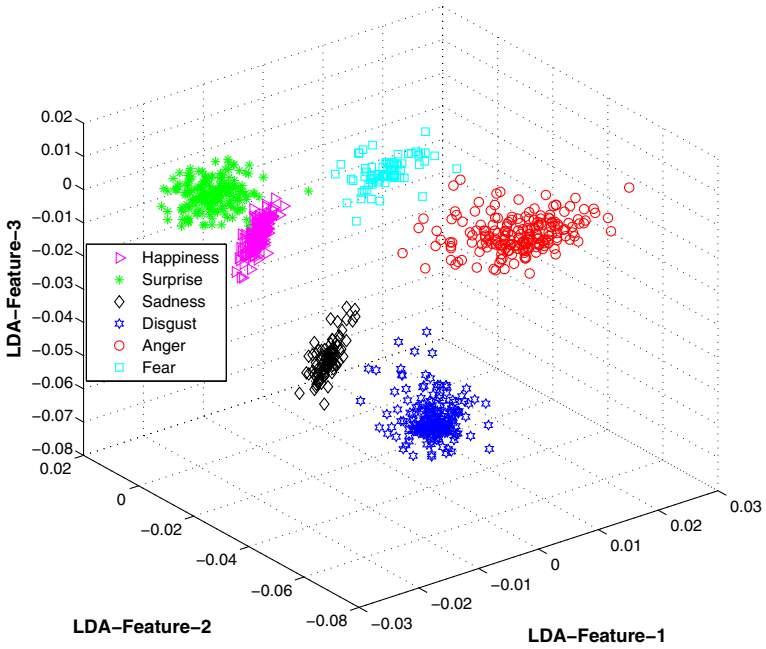

Fig. 9 3D feature plots of the CNF-FER for recognizing the expressions (on Cohn-Kanade dataset). It can be seen that the CNF-FER clearly classified the expressions classes

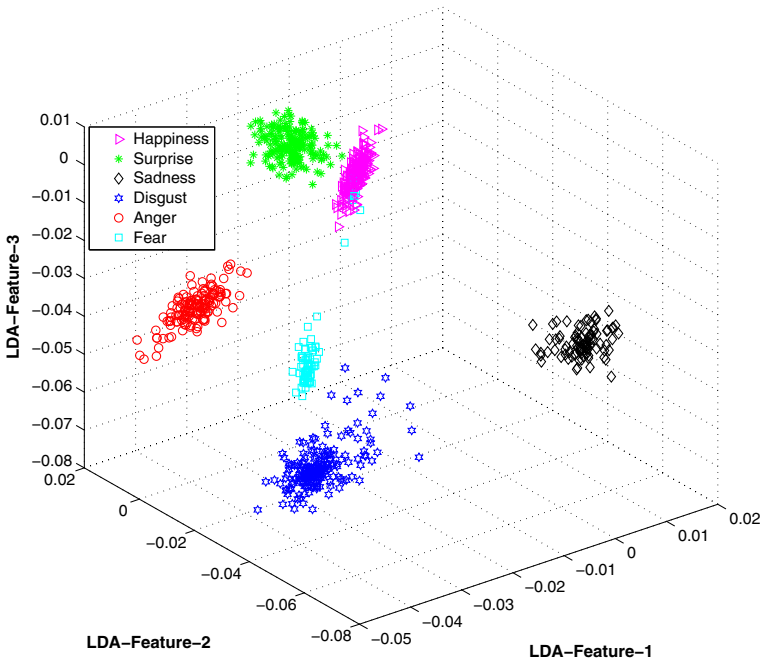

Fig. 10 3D feature plots of the CNF-FER for recognizing the expressions (on USTC-NVIE dataset). It can be seen that the CNF-FER clearly classified the expressions classes

Table 1 Confusion matrix of CF-FER using: (A) JAFFE dataset, (B) Yale B dataset, (C) Cohn-Kanade dataset, and (D) USTC-NVIE dataset of facial expressions (Unit: %)

	Happiness	Sadness	Anger	Disgust	Surprise	Fear
(A)						
Happiness	99	0	1	0	0	0
Sadness	0	100	0	0	0	0
Anger	0	1	99	0	0	0
Disgust	1	0	0	99	0	0
Surprise	0	1	0	0	98	1
Fear	1	0	0	0	0	99
Average			99.00			
(B)						
Happiness	100	0	0	0	0	0
Sadness	0	100	0	0	0	0
Anger	0	1	98	0	1	0
Disgust	0	1	0	99	0	0
Surprise	0	1	0	1	98	0
Fear	0	0	0	0	0	100
Average			99.17			
(C)						
Happiness	99	1	0	0	0	0
Sadness	1	98	0	1	0	0
Anger	0	0	100	0	0	0
Disgust	0	0	0	100	0	0
Surprise	1	0	0	1	98	0
Fear	0	0	0	0	0	100
Average			99.17			
(D)						
Happiness	99	0	0	0	1	0
Sadness	1	99	0	0	0	0
Anger	0	0	100	0	0	0
Disgust	1	0	0	99	0	0
Surprise	1	0	0	0	99	0
Fear	0	0	0	0	0	100
Average			99.33			

– **Recognition Rates of CNF-FER-based on Datasets**:
A set of experiments was performed in order to show the performance of CNF-FER system based on dataset. For these experiments, n-fold cross validation based on dataset was employed (in our case n=4), that means that out of four datasets, one dataset was

utilized as training data whereas the remaining three datasets were used as testing data, and this process was repeated four times, in which each data is used once for training and testing respectively. The weighted average recognition results of the CNF-FER systems on four datasets are shown in Table 2.

Table 2 Confusion matrix of CNF-FER system that is: (A) trained on JAFFE dataset and tested on Yale B, Cohn-Kanade, and USTC-NVIE datasets, (B) trained on Yale B face dataset and tested on JAFFE, Cohn-Kanade, and USTC-NVIE datasets, (C) trained on Cohn-Kanade dataset and tested on JAFFE, Yale B, and USTC-NVIE datasets, and (D) trained on USTC-NVIE dataset and tested on JAFFE, Yale B, and Cohn-Kanade datasets (Unit: %)

	Happiness	Sadness	Anger	Disgust	Surprise	Fear
(A)						
Happiness	87	2	3	4	3	1
Sadness	1	90	2	2	3	2
Anger	4	3	84	4	3	2
Disgust	1	2	3	88	3	3
Surprise	4	4	2	1	88	1
Fear	1	4	3	1	0	91
Average			88.00			
(B)						
Happiness	83	3	4	3	2	5
Sadness	1	90	2	2	3	2
Anger	2	3	87	3	3	2
Disgust	1	3	2	88	2	4
Surprise	3	3	4	3	84	3
Fear	2	4	2	3	4	85
Average			86.17			
(C)						
Happiness	87	3	2	3	1	4
Sadness	2	90	0	3	3	2
Anger	2	3	86	2	2	5
Disgust	1	3	2	88	3	3
Surprise	1	2	4	2	90	1
Fear	1	3	2	2	1	91
Average			88.67			
(D)						
Happiness	91	0	2	3	3	2
Sadness	2	85	3	4	3	3
Anger	1	3	89	2	4	1
Disgust	2	2	2	90	3	1
Surprise	4	3	1	2	88	2
Fear	2	2	3	2	1	90
Average			88.83			

Table 3 Confusion matrix of the CNF-FER using: (A) JAFFE dataset, (B) Yale B dataset, (C) Cohn-Kanade dataset, and (D) USTC-NVIE dataset, while removing the proposed feature selection method (Unit: %)

	Happiness	Sadness	Anger	Disgust	Surprise	Fear
(A)						
Happiness	92	2	1	1	3	1
Sadness	1	93	2	2	2	0
Anger	0	2	94	1	1	2
Disgust	0	5	0	95	0	0
Surprise	3	2	2	0	91	2
Fear	0	2	2	3	0	93
Average			93.00			
(B)						
Happiness	93	1	2	1	3	0
Sadness	1	94	0	2	1	2
Anger	2	3	92	3	0	0
Disgust	1	2	2	94	1	0
Surprise	3	0	2	0	95	0
Fear	1	3	0	3	93	
Average			93.50			
(C)						
Happiness	94	2	1	1	2	0
Sadness	0	95	2	1	1	1
Anger	1	1	96	0	1	1
Disgust	0	2	2	93	3	0
Surprise	0	1	1	3	95	0
Fear	2	1	1	2	0	94
Average			94.00			
(D)						
Happiness	94	2	1	1	2	0
Sadness	1	95	2	2	0	0
Anger	2	0	92	3	1	2
Disgust	0	2	3	93	2	0
Surprise	0	0	3	1	95	1
Fear	3	0	2	1	2	92
Average			93.50			

It can be seen from Table 2 that the CNF-FER does not achieve high recognition rate only on individual datasets, but also shows better performance when the system was trained on one dataset and tested on other datasets.

4.3 Third experiment

In order to assess the effectiveness of the proposed feature selection method, a series of sub-experiments were performed in this experiment. These experiments were performed using all of the four datasets and results are presented in Table 3.

Table 4 Comparison results of the proposed approaches with recent feature extraction methods (Unit: %)

Existing Works	[34]	[46]	[45]	[15]	[25]	[24]	CNF-FER
Average Accuracy Rate	92	86	87	85	96	94	99

It can be noted from Table 3 that the proposed feature selection method played a major role in the high recognition of CNF-FER. When we removed the proposed feature selection method, the recognition rate decreased signicantly. These results validate the problem of high similarity among the features of different expressions. The experimental results confirmed our analysis and provided clear evidence, allowing us to conclude that our proposed feature selection method selects a better feature set in terms of classification accuracy.

4.4 Fourth experiment

In the fourth experiment, the CNF-FER was compared against the following state-of-the-art FER methods: [15, 24, 25, 34, 45, 46]. JAFFE, Yale B, Cohn-Kanade, and USTC-NVIE datasets of facial expressions were used in this experiment. For a fair comparison, we borrowed the implementations of some of these methods, whereas for some methods, their published results are reported. The comparison has been performed under the same guidelines which were provided in their respective manuscripts. A 10-fold cross-validation scheme was utilized for each dataset (as described in Section 3). For the four datasets, the weighted average recognition rates of the existing methods and that of CNF-FER are presented in Table 4.

It is clear from Table 4 that the CNF-FER outperformed the existing state-of-the-art methods.

Moreover, in order to analyze the computational cost of the CNF-FER, we selected the most efficient method (that is, [34] from the above experiments of Table 4). The FER system of [34] took 1533 ms, 1498 ms, 2292 ms, and 1701 ms to recognize an expression frame from JAFFE, Yale B, Cohn-Kanade, and USTC-NVIE datasets of facial expressions, respectively. On the other hand, CNF-FER took 1290 ms, 1034 ms, 1908 ms, and 1865 ms to recognize an expression frame from the same datasets. Thus, the CNF-FER not only achieved high recognition rate, but it is also less expensive in terms of computational cost.

Moreover, the FER system of [34] has a complexity of $O(TQ^2M)$, where T is the length of an input sequence (i.e., expression frames), Q is the number of states, and M is the number of mixtures. On the other hand, the CNF-FER needs a maximum of $O(TM)$ to compute gradients. These experiments were performed in Matlab using an Intel Pentium Dual-Core^{TM} (2.5 GHz) with a RAM capacity of 3 GB.

5 Conclusion

A typical FER system consists of four modules: preprocessing, feature extraction, feature selection, and recognition. A great deal of research has been done for preprocessing, feature extraction, and recognition modules; however, feature selection is still an active research area.

In this paper, we have reviewed some recently developed algorithms for mutual information-based feature selection. We discussed the limitations of each method, and based

on our observations, we proposed our own method derived from the NMIFS with two improvements: the normalization of the mutual information and the feature-independent normalizing weights. For feature extraction, we have utilized the existing curvelet transform that has the capability to extract prominent features by keeping the line, curve, and edge information from each expression frame. The dimensions of the feature space were reduced by employing LDA. Finally, HMM was used as the recognizer.

The CNF-FER was tested and validated using four publicly available standard datasets. For each dataset, 10-fold cross-validation scheme was employed. The CNF-FER achieved a weighted average recognition accuracy of 99 %, which is a significant improvement over the recognition rates of existing FER systems. Moreover, from computational perspective, the CNF-FER is less expensive than existing methods.

The performance of CNF-FER is yet to be investigated in real-time, because there exist several factors in real-time environment that might decrease the performance of CNF-FER, such as background clutter, image rotation and blur, and varying face angles. Therefore, further study is needed to tackle these issues and maintain the same high recognition rate in real environment.

Acknowledgments This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) (No. 2013-067321)).

This research was also supported by the MSIP (Ministry of Science, ICT & Future Planning), Korea, under the ITRC (Information Technology Research Center) support program supervised by the NIPA (National IT Industry Promotion Agency) (NIPA-2014-(H0301-14-1003).

References

1. Abd El Meguid M, Levine M (2014) Fully automated recognition of spontaneous facial expressions in videos using random forest classifiers. IEEE Trans Affect Comput 5(2):141–154
2. Battiti R (1994) Using mutual information for selecting features in supervised neural net learning. IEEE Trans Neural Netw 5(4):537–550
3. Baudat G, Anouar F (2000) Generalized discriminant analysis using a kernel approach. Neural Comput 12(10):2385–2404
4. Baum LE (1972) An equality and associated maximization technique in statistical estimation for probabilistic functions of markov processes. Inequalities 3:1–8
5. Belhumeur PN, Hespanha JP, Kriegman D (1997) Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. IEEE Trans Pattern Anal Mach Intell 19(7):711–720
6. Bettadapura V (2012) Face expression recognition and analysis: the state of the art. arXiv preprint. arXiv:12036722
7. Candes E, Demanet L, Donoho D, Ying L (2006) Fast discrete curvelet transforms. Multiscale Modeling & Simulation 5(3):861–899
8. Candes EJ, Donoho DL (2000) Curvelets, multiresolution representation, and scaling laws. In Proc. SPIE, vol 4119, pp 1–12
9. Dash M, Liu H (1997) Feature selection for classification. Intell Data Anal 1(3):131–156
10. Estevez PA, Tesmer M, Perez CA, Zurada JM (2009) Normalized mutual information feature selection. IEEE Trans Neural Netw 20(2):189–201
11. Fasel B, Luettin J (2003) Automatic facial expression analysis: a survey. Pattern Recogn 36(1):259–275
12. Filko D, Martinović G (2013) Emotion recognition system by a neural network based facial expression analysis. Automatika: Journal Control Meas Electron Comput Commun 54(2):263–272
13. Fodor IK (2002) A survey of dimension reduction techniques
14. Georghiades AS, Belhumeur PN, Kriegman D (2001) From few to many: Illumination cone models for face recognition under variable lighting and pose. IEEE Trans Pattern Anal Mach Intell 23(6):643–660
15. Ghimire D, Lee J (2013) Geometric feature-based facial expression recognition in image sequences using multi-class adaboost and support vector machines. Sensors 13(6):7714–7734
16. Hall MA (1999) Correlation-based feature selection for machine learning. PhD thesis. The University of Waikato

17. He S, Wang S, Lv Y (2011) Spontaneous facial expression recognition based on feature point tracking. In 2011 Sixth International Conference on Image and Graphics (ICIG). IEEE, pp 760–765
18. Hossain MA, Sanyal G (2012) A new improved tactic to extract facial expressions based on genetic algorithm and wvdf. Int J Adv Inf Technol 2(5):37–44
19. Kamimura R (2011) Structural enhanced information and its application to improved visualization of self-organizing maps. Appl Intell 34(1):102–115
20. Kanade T, Cohn JF, Tian Y (2000) Comprehensive database for facial expression analysis. In Proceedings. Fourth IEEE International Conference onAutomatic Face and Gesture Recognition, 2000. IEEE, pp 46–53
21. Kashyap KL, Shantaiya S (2012) Study and analysis of statistical features of face expression in noisy environment. Int J Image Process Vis Sci 1(2):29–34
22. Kotsia I, Pitas I (2007) Facial expression recognition in image sequences using geometric deformation features and support vector machines. IEEE Trans Image Process 16(1):172–187
23. Kwak N, Choi CH (2002) Input feature selection for classification problems. IEEE Trans Neural Netw 13(1):143–159
24. Liu M, Shan S, Wang R, Chen X (2014a) Learning expressionlets on spatio-temporal manifold for dynamic facial expression recognition. In IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2014. CVPR 2014
25. Liu P, Han S, Meng Z, Tong Y (2014b) Facial expression recognition via a boosted deep belief network. In IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2014, CVPR 2014
26. Lyons M, Akamatsu S, Kamachi M, Gyoba J (1998) Coding facial expressions with gabor wavelets. In Proceedings. Third IEEE International Conference on Automatic Face and Gesture Recognition, 1998. IEEE, pp 200–205
27. Mika S (2002) Kernel fisher discriminants. PhD thesis, Universitätsbibliothek
28. Mika S, Ratsch G, Weston J, Scholkopf B, Mullers K (1999) Fisher discriminant analysis with kernels. In Proceedings of the 1999 IEEE Signal Processing Society Workshop. Neural Networks for Signal Processing IX, 1999. IEEE, pp 41–48
29. Pagariya RR, Bartere MM (2013) Facial emotion recognition in videos using hmm 3(4):111–118
30. Peng H, Long F, Ding C (2005) Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans Pattern Anal Mach Intell 27(8): 1226–1238
31. Piatkowska E, Martyna J (147) Spontaneous facial expression recognition: automatic aggression detection. In Hybrid Artificial Intelligent Systems. Springer
32. Rabiner L (1989) A tutorial on hidden markov models and selected applications in speech recognition. Proc IEEE 77(2):257–286
33. Ramirez-Gutierrez K, Sanchez-Perez D, Perez-Meana H, Fujita H, Sasaki J (2010) Face recognition and verification using histogram equalization Selected Topics in Applied Computer Science. WSEAS, pp 85–89
34. Ramirez Rivera A, Castillo R, Chae O (2013) Local directional number pattern for face analysis: Face and expression recognition. IEEE Trans Image Process 22(5):1740–1752
35. Russell JA (2003) Core affect and the psychological construction of emotion. Psychol Rev 110(1):145
36. Samaria FS (1994) Face recognition using hidden markov models. PhD thesis. University of Cambridge
37. Schels M, Schwenker F (2010) A multiple classifier system approach for facial expressions in image sequences utilizing gmm supervectors. In 2010 20th International Conference on Pattern Recognition (ICPR). IEEE, pp 4251–4254
38. Shan C, Gong S, McOwan PW (2009) Facial expression recognition based on local binary patterns: A comprehensive study. Image Vis Comput 27(6):803–816
39. Siddiqi MH, Lee S, Lee YK, Khan AM, Truc PTH (2013) Hierarchical recognition scheme for human facial expression recognition systems. Sensors 13(12):16682–16713
40. Siddiqi MH, Ali R, Khan AM, Kim ES, Kim GJ, Lee S (2014) Facial expression recognition using active contour-based face detection, facial movement-based feature extraction, and non-linear feature selection. Multimedia Systems
41. Smith BM, Brandt J, Lin Z, Zhang L (2014) Nonparametric context modeling of local appearance for pose-and expression-robust facial landmark localization
42. Tang M, Chen F (2013) Facial expression recognition and its application based on curvelet transform and pso-svm. Optik-Int J Light Electron Optics 124(22):5401–5406
43. Uddin MZ, Lee J, Kim TS (2009) An enhanced independent component-based human facial expression recognition from video. IEEE Trans Consum Electron 55(4):2216–2224

44. Wang S, Liu Z, Lv S, Lv Y, Wu G, Peng P, Chen F, Wang X (2010) A natural visible and infrared facial expression database for expression recognition and emotion inference. IEEE Trans Multimedia 12(7):682–691
45. Zhang S, Zhao X, Lei B (2012) Robust facial expression recognition via compressive sensing. Sensors 12(3):3747–3761
46. Zhao X, Zhang S (2011) Facial expression recognition based on local binary patterns and kernel discriminant isomap. Sensors 11(10):9573–9588

Muhammad Hameed Siddiqi is PhD student in Ubiquitous Computing (UC) Lab, Department of Computer Engineering, Kyung Hee University, South Korea. He did his Master of Engineering Degree from Department of Computer Engineering, Kyung Hee University, South Korea in 2012, and Bachelor of Computer Science (Hons) from Islamia College university of Peshawar, N-W.F.P, Pakistan in 2007. He was a Graduate Assistant at Universiti Teknologi PETRONAS, Malaysia from 2008 to 2009. His research interest is Image Processing, Pattern Recognition, Machine Intelligence and Activity Recognition.

Rahman Ali is a PhD student in Ubiquitous Computing Laboratory (UCLab), Department of Computer Engineering, Kyung Hee University, South Korea. He got his M.Phil Degree in Computer Science from Department of Computer Science, University of Peshawar, Pakistan in 2009. He secured his M.Sc Degree (in Computer Science) from Hazara University, Mansehra, Pakistan in 2005 and Bachelor Degree (in Computer Science) from Govt. Jahanzeb College Saidu Sharif, Swat, Pakistan back in 2002. He has been a lecturer in Computer Science at University of Peshawar since 2009. He also served Institute of Information Technology, University of Science and Technology, Bannu as a lecturer in computer science and the Laboratoire d'Informatique de l'Universit du Maine, France as a research assistant. His current research interest is Machine Learning, Knowledge Acquisition and Reasoning.

Muhammad Idris is pursuing Masters leading to Ph.D in Computer engineering at Kyung Hee University, Republic of Korea. He has been Research Assistant at KTH-AIS laboratory, SEECS- NUST Pakistan. His current research interests include Big Data Analytics, Distributed Computing and Security

Adil Mehmood Khan received his Ph.D. degree from the Department of Computer Engineering of Kyung Hee University, Republic of Korea in 2011. He is now working as a faculty member with the Division of Information and Computer Engineering, Ajou University, Republic of Korea. His research interest includes pattern recognition, signal processing, ubiquitous computing, and machine learning.

Eun Soo Kim is Professor of Dept. of Electronics Eng. At the Kwangwoon University in Seoul Korea. He received his Ph.D in Electronics from Yonsei University, Seoul Korea. He was a Visiting Professor at the Dept. of Electrical Eng., California Institute of Technology, USA. During 1987-1988. He served the President of the Society of 3D Broadcasting and Imaging, the President of the 3D Fusion Industry Consortium, and the President of the Korean Information and Communications Society during 2000-2011. In addition, he has been honored to serve the Editor-in-Chief of '3D Research' (www.springer.com, www.3dresearch.org) since 2010, and the General-chair of the International meeting of 'Collaborative Conference on 3D & Materials Research (CC3DMR) since 2011'. He has also organized and annually host the 3D event of 'International 3D Fair'in collaboration with '3D Consortium of Japan', China 3D Industry Association of China and 3D@Home Consortium of USA since 2006. Now, he has more than 100 domestic and overseas patents, and he published more than 400 papers in the international journals and conferences. He receive the Order of Science and Technology Merit from the President of the Republic of Korea, Roh Moo-hyun in 2003, and the Order of Service Merit from the President of the Republic of Korea, Lee Myung-Bak in 2012.

Min Cheol Whang received the M.S. and Ph.D. degrees in Biomedical Engineering from the Georgia Institute of Technology, Atlanta, Georgia, USA, in 1990 and 1994, respectively. He has been a Professor in the Division of Digital Media Technology and Department of Emotion Engineering, Graduate School, Sangmyung University, Seoul, Korea since March 1998. His research interests include Human Computer Interaction, Emotion Engineering, Human Factors and Bioengineering.

Sungyoung Lee received his B.S. from Korea University, Seoul, Korea. He got his M.S. and Ph.D. degrees in Computer Science from Illinois Institute of Technology (IIT), Chicago, USA in 1987 and 1991 respectively. He has been a professor in the department of Computer Engineering, Kyung Hee University, Korea since 1993. He is a founding director of the Ubiquitous Computing Laboratory, and has been affiliated with a director of Neo Medical ubiquitous-Life Care Information Technology Research Center, Kyung Hee University since 2006. Before joining Kyung Hee University, he was an assistant professor in the Department of Computer Science, Governors State University, llinois, USA from 1992 to 1993. His current research focuses on Ubiquitous Computing and Applications, Wireless Ad-hoc and Sensor Networks, Context-aware Middleware, Sensor Operating Systems, Real-Time Systems and Embedded Systems, Activity and Emotion Recognition. He is a member of ACM and IEEE.