Facial Expression Recognition Based on Significant Face Components Using Steerable Pyramid Transform

Ayşegül UÇAR
Frat University, Mechatronics Engineering Department, Elazığ, Turkey, e-mail: agulucar@ieee.org

Abstract – Facial expression recognition is a challenging problem in many areas such as computer vision and human-computer interaction. To extract an effective facial features and then to classify them are the best important points of facial expression recognition process. In this article, a new automatic facial expression recognition algorithm is proposed in order to further enhance the recognition performance in terms of these two points. First, it is detected the some specific components of face, such as the mouth, eyes, eyebrows, and nose by Viola-Jones algorithm. Secondly it is extracted features by applying local Steerable Pyramid Transform (SPT) to each of facial component images. Thirdly it is used Support Vector Machines (SVM) classifiers for facial expression verification. Finally the classifier outputs are combined by decision fusion. The experiments on the Japanese Female Facial Expression (JAFFE) database and the Cohn-Kanade database show that the proposed Component based - Facial Expression Recognition (CFER) algorithm improves facial expression recognition performance compared to an algorithm combining SPT and Principal Component Analysis (PCA) using whole face images to the results in the literature.

Keywords: Facial Expression Recognition, Steerable Pyramid Transform, SVM.

1 Introduction

Facial expression recognition is currently an active research topic and challenging due to its important in several real-world applications such as border security, forensic, virtual reality, and robotics in areas of human-computer interaction and computer vision [1-4]. Recently considerable research efforts relating to especially the facial expression recognition with high correctness have been increased the applications of facial expression recognition. In the Facial Action Coding System (FACS) in [5], the basic facial expressions were semantically coded as happiness, sadness, fear, anger, disgust, and surprise. However the expressions present a wide variation with respect to the individuals. Hence to assign to one to from six basic categories a facial expression relating to a person and to recognize that person from the images including different facial expressions are challenging problem [6-7].

A number of facial expression algorithms have been proposed. They could be classified into two categories: 1) Feature-based methods use the shape and location information of significant regions appearing the face geometry as the feature vector [8-9]. These methods provide the good intuitive results with high performance thanks to physical points but they need to accurate and reliable detection of face points. 2) Appearance-based methods extract facial features by applying a set of filters to whole face or a part of face and obtain high dimension data. The methods are firstly followed by some dimension reducing algorithms such as Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), and locality preserving and then the classifying algorithms such as Support Vector Machines (SVMs) and neural networks [10-11]. Appearance-based methods are capable of representing with a powerful subspace of overall space but they require large memory usage with respect to feature-based ones.

In recent years, a lot of methods have been successfully applied for facial expression recognition, including optical flow [12-13], Gabor wavelets [14-15], Local Binary Patterns [11], Zernike moments [6], and image ratio features [16]. The above studies have presented that there have still need to many researches in order to improve facial expression recognition performance. In this paper, Steerable Pyramid Transform (SPT) algorithm is used as an effective feature extractor in facial expression recognition.

The aim of this paper is to partially combine the advantages of feature based methods and appearance-based methods. In this paper, the Component based - Facial Expression Recognition (CFER) algorithm is proposed. The CFER algorithm consists of the following steps: i) Viola Jones algorithm is used to detect the face in the images and then the specific component regions relating to the cropped face image, ii) the each component image is preprocessed in order to provide more illumination invariant, iii) SPT algorithm is applied for extracting the local features relating to each facial component image, iv) the SVM classifiers are separately applied to each component image, v) the classifier results are combined.

Feature extraction step and classifying step are of great important in facial expression recognition. In this paper, the feature set is extracted firstly by applying SPT to the partitioned specific regions and then by computing some statistical features relating to each local portion. SVM classifiers that are one of the best classifiers in the literature are used to classify facial expressions. When the proposed component based algorithm is compared to one using whole image, it presents more robustness to illumination and orientation and has higher recognition performance. In addition, since physical points are detected and less feature numbers are determined, the proposed CFER algorithm does not need to any dimension reduction method. Moreover, CFER can be used in the forensic applications concerning occluded or partial face images.

The remainder of this paper is organized as follows. In Section 2, the SPT is introduced. In Section 3, it is given a
short review about SVMs. In Section 4, the proposed CFER algorithm is described. In Section 5, the results from experiments are comparatively discussed with results in the literature. In Section 6, the conclusions are summarized.

2 Steerable Pyramid Transform

The Steerable Pyramid is a linear multi-resolution image decomposition method [17]. In SPT, an image is subdivided into a collection of subbands localized at different scales and orientations. The transform is invariant from translation and rotation.

Fig. 1 shows the operating logic of SPT. Input image is initially decomposed into the subbands of high-pass and low-pass using a high-pass filter $H_0(w)$ and a band lowpass filter $L_0(w)$. The lowpass subband is then decomposed into $K$-oriented portions using the bandpass filters $B_k(w)$ ($k=0,1,...,K-1$) and into a lowpass subband $L_1$ [17]. The process is done recursively by down and up subsampling the lower lowpass subband by a multiplier of 2 along the rows and columns.

![Fig. 1 Block diagram of pyramid decomposition [12]](image)

Fourier domain formulations of the lowpass filters and highpass filters are defined as:

$$L_0(r, \theta) = L \left( \frac{r}{2}, \theta \right),$$  \hspace{1cm} (1)

$$H_0(r, \theta) = H \left( \frac{r}{2}, \theta \right),$$  \hspace{1cm} (2)

where $r$ and $\theta$ are the polar frequency coordinates.

$$L(r, \theta) = \begin{cases} 2 \cos \left( \frac{\pi}{2} \log_2 \left( \frac{2r}{\pi} \right) \right), & \frac{\pi}{4} < r < \frac{\pi}{2} \\ \frac{1}{2}, & r \geq \frac{\pi}{4} \\ 0, & r \leq \frac{\pi}{4} \end{cases}$$  \hspace{1cm} (3)

$$B_k(r, \theta) = H(r)G_k(\theta), \quad k \in [0, K - 1]$$  \hspace{1cm} (4)

where $B_k(r, \theta)$ is the $K$ directional bandpass filters used in iterative steps with radial and angular parts.

$$H(r) = \begin{cases} \cos \left( \frac{\pi}{2} \log_2 \left( \frac{2r}{\pi} \right) \right), & \frac{\pi}{4} < r < \frac{\pi}{2} \\ 1, & r \geq \frac{\pi}{4} \\ 0, & r \leq \frac{\pi}{4} \end{cases}$$  \hspace{1cm} (5)

$$G_k(\theta) = \frac{(K-1)!}{\sqrt{2(2\pi)^{K-1}}} \left( 2 \cos \left( \theta - \frac{\pi k}{K} \right) \right)^{K-1}.$$  \hspace{1cm} (6)

Fig. 2 shows all filtered images at 3 scales consisting of 128x128, 64x64, and 32x32 and 4 orientation subbands consisting of -\pi /4, 0, \pi /4, and \pi /2 on a cropped original image of Cohn-Kanade database.

![Fig. 2 Steerable pyramid transform (K=4 and J=3) on a cropped original image of Cohn-Kanade database](image)

3 A Short Review for Support Vector Machines

For a training set of $L$ points of the form $(x^1, y^1), ..., (x^L, y^L), x \in \mathbb{R}^n, y \in \{-1, 1\}$, SVMs define a nonlinear mapping $\phi(x)$ from input space to a higher dimensional feature space. When the input data is mapped to feature space, the two training classes may linearly separable in the feature space by the optimal discriminant function defined by

$$\ell(x) = w^T \phi(x) + b$$  \hspace{1cm} (7)

where $w \in \mathbb{R}^n$ is a weight vector and $b \in \mathbb{R}$ is a bias [18-19]. The optimal discriminant function is determined by using maximizing the margin and by minimizing the misclassification error. The margin is defined as the distance between the discriminant function and the samples nearest to the hyperplane [19]. SVM classifier solves the following primal optimization:

$$\min_{w,b,\xi} L_{primal}(w, \xi) = C \sum_{i=1}^L \xi_i + \frac{1}{2} \|w\|^2$$  \hspace{1cm} (8)

subject to $y_i[w^T \phi(x^i) + b] \geq 1 - \xi_i$

The first term in Eq. (8) penalizes the sum of absolute errors $\xi_i$. The slack variables allow some points to be
misclassified. The second term compels the margin to maximum. C is a trade-off constant between maximum margin and classification error. A higher C value provides a larger penalty for classification error.

By eliminating all primal variables in Eq. (8) and introducing Lagrange multipliers \( \lambda_i \)'s, the dual quadratic optimization problem is constructed as the following one

\[
\max_{\lambda} L_{\text{dual}}(\lambda) = -\frac{1}{2} \sum_{i,j=1}^{L} \lambda_i \lambda_j y_i y_j K(x_i, x_j) + \sum_{i=1}^{L} \lambda_i \tag{9}
\]

subject to \( \sum_{i=1}^{L} y_i \lambda_i = 0, \quad 0 \leq \lambda_i \leq C, \quad i = 1, \ldots, L \)

where \( K(x_i, x_j) \) is called as the kernel function in the form of inner product of \( \varphi(x_i)^T \varphi(x_j) \).

The discriminant function is easily defined

\[
l(x) = \text{sign} \left( \sum_{i \in \mathcal{S}} y^i \lambda_i K(x_i, x^j) + b \right) \tag{10}
\]

where the Support Vectors (SVs) are defined as the data points \( x^j \) corresponding to \( \lambda_i > 0 \).

4 Proposed Facial Expression Recognition Algorithm

In this section, it is presented the proposed CFER algorithm. The block schema of the proposed CFER algorithm is given in Fig. 3. The CFER algorithm is applied at seven folds:

1. Detected the face region in image by using Viola Jones algorithm,
2. Preprocessing each image by histogram equalization in order to provide further illumination invariance [20],
3. Determined the two face regions relating to eye pairs and nose by using Viola Jones algorithm [21] and extended the regions to cover the regions of eyebrow and mouth,
4. Selected the significant face regions such as mount, nose, eye, and eyebrow on the obtained region in previous stage. The selection is standardized for all images without any manual process. Each specific region is partitioned to local regions [22],
5. Extracted the representative feature set by applying SPT to each face component image,
6. Apply SVM classifiers to the qualified coefficients,
7. Verify the facial expression recognition by a weighted majority voting rule to the classifier output.

Fig. 3. The block schema of the proposed CFER algorithm

5 Experimental Results

To validate the accuracy of the proposed algorithm, it was used the Japanese Female Facial Expression database (JAFFE) [23] and the Cohn-Kanade database [24].

The JAFFE database consists of 213 grayscale facial expression images obtained from 10 female [23]. There are seven different facial expressions, such as happy, angry, disgust, fear, sad, and surprise in addition to one neutral face expression. Each subject has two to four different images for each expression. Each image has the resolution of 256 x 256 pixels. Fig. 4. shows some images comprising seven basic facial expressions from the JAFFE database.

Cohn-Kanade database is one of the most comprehensive benchmarks for facial expression databases [24]. The database consists of 3911 images from 138 subjects ranged in age from 18 to 30 years. Sixty five percent is female, 15 percent is African-American and 3 percent is Asian or Latino. Fig. 5. shows some images comprising seven basic facial expressions from the Cohn-Kanade database. All image sequences have the resolution of 640 x 480 or 640 x 490 pixels.
In this paper, all the images of 10 basic expressions from the JAFFE database are used. Seven class classification problem was constructed including neutral expression. For the Cohn-Kanade database, 258 of the image sequences given in [25] were selected such that each image sequence has eight images. Eight images were chosen from the last peak images of each image sequence. The image sequences consist of 2020 images (176 Anger, 264 Disgust, 256 Fear, 624 Happy, 324 Sadness, and 376 Surprise). To evaluate the generalization performance to all subjects in the experiments, it was adopted a 10-fold cross-validation testing scheme. It was partitioned the dataset randomly into ten groups of roughly equal numbers of subjects. Nine groups were used as the training data to train classifiers, while the remaining group was used as the test data. The above process was repeated ten times for each group in turn to be omitted from the training process. It was reported the average recognition results on the test sets.

The images of 10 subjects in the JAFFE database are classified into 10 sets, each of which includes images of one subject. Similarly, all images in the Cohn-Kanade database are classified into 10 similar sets and all images of one subject are included in the same set.

In the first stage of the CFER algorithm, the faces in the images were detected by using Viola Jones algorithm. Secondly all the images were cropped with respect to the determined face locations and scaled to 128×128 pixels resolution. Thirdly the eye and nose regions of face were detected by Viola Jones algorithm. The left hand side of Fig. 6 shows these two regions. The specific regions relating to eyebrow, eye, nose, and mouth were then determined. The regions relating to eyebrow, eye, nose, and mouth have resolutions 20×100, 15×100, 25×60, and 40×60, respectively. This pixel resolution was generalized for all databases. The right hand side of Fig. 6 shows these four regions. All component images were decomposed using SPT at 3 scales (128×128, 64×64, and 32×32) and 4 orientation subbands (0, π/4, 0, π/4, and π/2) for each specific component. Local regions were constructed by an efficient pixel number of 5×5.

In experiments, only all subbands relating to the first scale were used since the results of the first scale were better than the others. Three statistical features such as mean, entropy, and variance of each local image part were extracted. It was extracted the regions relating to eyebrow, eye, nose, and mouth are extracted the features of 60, 30, 36, and 72, respectively. Since the feature number was not large, there was no need to any additional dimension reduction algorithm. Fig. 7 shows the best important steps applying local SPT of CFER algorithm on a Cohn Kanade subject.

In this paper, it was used SVM classifiers in [18]. All features were classified by four SVM classifiers. The kernel and regularization parameters of SVM are determined by 10-fold cross validation scanning the chosen parameter range. The obtained results are fused by the weighed majority voting rule. The weights of constructed classifiers for each subband are determined by dividing itself correctness to the obtained total correctness.
components by PCA. All results in Tables 1 and 2 show that the performance can be increased up to around at least 5% by using CFER algorithm compared to the other. Table 3 shows that the proposed approach outperforms all of seven benchmarked approaches in [11, 15, 25-29] when both the JAFFE database and Cohn-Kanade database. In the literature, the selection of the training and testing sets and the number of facial expressions presents a wide of variation. For example [27] uses every two images from the peak frame, [11] gives

the results based on five-fold cross validations and five expressions, and [15] uses leave-one-subject-out with seven facial expressions. It is difficult to make a right comparison with them. However the results of this paper appear satisfying.

<table>
<thead>
<tr>
<th>Regions</th>
<th>Anger</th>
<th>Happy</th>
<th>Sadness</th>
<th>Disgust</th>
<th>Surprise</th>
<th>Fear</th>
<th>Neutral</th>
</tr>
</thead>
<tbody>
<tr>
<td>Eyebrows</td>
<td>87.45</td>
<td>96.00</td>
<td>89.76</td>
<td>91.11</td>
<td>95.09</td>
<td>92.78</td>
<td>90.89</td>
</tr>
<tr>
<td>Eyes</td>
<td>88.88</td>
<td>92.24</td>
<td>91.00</td>
<td>91.23</td>
<td>89.05</td>
<td>82.22</td>
<td>86.90</td>
</tr>
<tr>
<td>Nose</td>
<td>89.12</td>
<td>88.80</td>
<td>78.00</td>
<td>88.23</td>
<td>91.92</td>
<td>98.12</td>
<td>96.56</td>
</tr>
<tr>
<td>Mouth</td>
<td>89.43</td>
<td>100</td>
<td>90.13</td>
<td>92.12</td>
<td>94.01</td>
<td>94.67</td>
<td>95.80</td>
</tr>
<tr>
<td>Fused</td>
<td>91.70</td>
<td>100</td>
<td>91.75</td>
<td>95.67</td>
<td>98.54</td>
<td>100</td>
<td>97.80</td>
</tr>
<tr>
<td>Whole Image</td>
<td>83.34</td>
<td>95.12</td>
<td>89.56</td>
<td>90.43</td>
<td>88.12</td>
<td>93.34</td>
<td>92.23</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Regions</th>
<th>Anger</th>
<th>Happy</th>
<th>Sadness</th>
<th>Disgust</th>
<th>Surprise</th>
<th>Fear</th>
<th>Neutral</th>
</tr>
</thead>
<tbody>
<tr>
<td>Eyebrows</td>
<td>90.32</td>
<td>92.02</td>
<td>88.56</td>
<td>90.45</td>
<td>89.56</td>
<td>90.56</td>
<td></td>
</tr>
<tr>
<td>Eyes</td>
<td>82.54</td>
<td>98.01</td>
<td>87.56</td>
<td>89.23</td>
<td>94.75</td>
<td>89.80</td>
<td></td>
</tr>
<tr>
<td>Nose</td>
<td>88.54</td>
<td>95.00</td>
<td>89.02</td>
<td>91.12</td>
<td>94.75</td>
<td>91.54</td>
<td></td>
</tr>
<tr>
<td>Mouth</td>
<td>88.75</td>
<td>99.56</td>
<td>86.75</td>
<td>90.12</td>
<td>94.75</td>
<td>92.02</td>
<td></td>
</tr>
<tr>
<td>Fused</td>
<td>92.45</td>
<td>100</td>
<td>89.89</td>
<td>92.76</td>
<td>96.67</td>
<td>94.03</td>
<td></td>
</tr>
<tr>
<td>Whole Image</td>
<td>85.45</td>
<td>92.23</td>
<td>81.75</td>
<td>79.56</td>
<td>89.43</td>
<td>89.01</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>JAFFE</th>
<th>Cohn-Kanade</th>
</tr>
</thead>
<tbody>
<tr>
<td>CFER</td>
<td>95.42</td>
<td>96.22</td>
</tr>
<tr>
<td>[26]</td>
<td>88.13</td>
<td>-</td>
</tr>
<tr>
<td>[27]</td>
<td>92.93</td>
<td>94.48</td>
</tr>
<tr>
<td>[28]</td>
<td>91.00</td>
<td>-</td>
</tr>
<tr>
<td>[29]</td>
<td>83.84</td>
<td>95.87</td>
</tr>
<tr>
<td>[25]</td>
<td>89.13</td>
<td></td>
</tr>
<tr>
<td>[11]</td>
<td>81.00</td>
<td>95.10</td>
</tr>
<tr>
<td>[15]</td>
<td>89.50</td>
<td>91.51</td>
</tr>
</tbody>
</table>

### 6 Conclusions

This paper presents a new facial expression recognition based on local SPT and histogram equalization. The important steps of this algorithm are to determine the specific facial regions and to compute the statistical features such as entropy, mean, and variance using local SPT and then to fuse the outputs of SVM classifiers applying to the features of each local patch. The recognition performances obtained on JAFFE and Cohn-Kanade databases have been shown to significantly better those in the literature and local SPT algorithm applied to whole face. Furthermore, the results shows that the proposed algorithm is robust to occlusion or missing information thanks to component based representations. The feature plan is to prove that the algorithm has the pose invariant property on the different datasets and to extent it to 3D facial expression.

**Acknowledgements:**

This paper was supported by the Firat University Scientific Research Projects Foundation (No. MF.12.33).

### 7 References

2. I. Cohn, N. Sebe, L. Chen, A. Garg, and T. S. Huang "Facial expression recognition from video sequences:


