A Comparative Study of Facial Emotion Classification

Tianming Hu, L.C. De Silva
Dept. Electrical Engineering, National University of Singapore
10 Kent Ridge Crescent, Singapore 119260

Abstract

Face plays an important role in communication so automatic recognition of facial emotion is an important addition to computer vision research. This paper presents a comparative study of two types of approaches to facial emotion classification on single images. Gabor wavelets is the technique we employ here to extract features from upper face and lower face. Linear discriminant function (LDF) is applied first for classification. In this type, we compare principle component analysis (PCA) and Fisher linear discriminant (FLD). In the second type of neural network, we focus on multi-layer perceptron (MLP), where single big MLP and multiple MLP classifier are compared. The experimental results show PCA outperforms FLD, and multiple MLPs classifier beats single big MLP. In addition, it is also indicated that Gabor coefficients at high frequency and vertical orientation may contain more information about facial emotion.

1. Introduction

Facial emotion recognition plays an important role in many applications. For instance, in man machine interface, machines will be more friendly and smarter if it can perceive the emotion of the user. In video conference where low band width is desirable, the facial image data can be reduced to some parameters of expression if the real image can be encoded at one end to some parameters and decoded at the receiver end. Other fields where facial emotion can provide information include basic research on the brain, education, criminal justice, medicine, international relationships and so on.

Approaches to facial emotion classification can be divided into three categories: static, semi-static and dynamic. In static approaches classification of a facial emotion is performed using a single face image. These approaches are mainly based on the Facial Action Coding System (FACS), which is developed in order to allow human psychologists to code emotion from still images. Two types of feature extraction techniques are commonly employed in this category: geometry based and wavelet transform based. Semi-static approaches extract temporal information by using only two frames. One represents the face in its neutral condition and one shows the face at the peak of a certain emotion. By comparing the two frames, motion feature is extracted. Dynamic approaches are applied to video image sequence by tracking emotion development.Often optical flow is used in this category because of its inherent property to catch motion and Hidden Markov model (HMM) has been used to model the sequence.

We study five emotions in our database: neutral, anger, dislike, joy and surprise(see Fig. 1). Section [2] gives a brief review of Gabor wavelets as a technique to extract features from facial images. In the sections to follow, LFD and NN are discussed in section [3] and section [4] respectively. Experimental results are given in section [5] and a short conclusion is given in section [6].

Figure 1: five emotions

2. Feature Extraction

Gabor wavelets are a relatively new technique to encode facial emotion and facial action unit (AU) [4, 5]. With a set of multi-frequency, multi-orientation Gabor wavelets, a face image can be transformed by computing the coefficients at certain key points such as brow corner or over the entire face. Past study shows Gabor wavelets are more powerful than other techniques in encoding face. Zhang *et al.* [5]compared two kinds of features, one being geometric position of a set of fiducial points on a face, and the other being Gabor coefficients at them. Using a 3 layer MLP

to classify the feature into one of the basic six emotions, the recognition rate on the test set was around 60% for the geometric position and around 90% for the Gabor coefficients. Donato *et al.* [4] explored and compared many techniques, including optical flow and Gabor representation, for recognizing facial actions in sequences of images. The recognition rate with Gabor representation was around 95%, much better than the result with optical flow, 85%. Another advantage of Gabor wavelets is that it is robust to small changes in image registration caused by say, slight rigid motion of the head, while other methods such as optical flow do not have this robust property.

For each gray image of 384×288 pixels in our database, centers of eyes and mouth are located manually. Then, crop two windows containing the regions of interest, upper face and lower face respectively. The upper face window is rescaled to size 65×30 pixels and the lower face window to 50×30 pixels. The location of centers of eyes and mouth coincide approximately over all images. Evenly divide both upper and lower face window into small grid of 5×5 pixels. Compute Gabor coefficients at the center of each grid as follows. Given an image represented by $I(\mathbf{x})$ where $\mathbf{x} = (x, y)$, G_i is defined as a convolution

$$G_i(\mathbf{x'}) = \int I(\mathbf{x})\psi_i(\mathbf{x} - \mathbf{x'})d^2\mathbf{x} \tag{1}$$

with a family of Gabor kernels ψ_i

$$\psi_i(\mathbf{x}) = \frac{|\mathbf{k}_i|^2}{\sigma^2} e^{-\frac{|\mathbf{k}_i|^2|\mathbf{x}|^2}{2\sigma^2}} [e^{j\mathbf{k}_i\mathbf{x}} - e^{-\frac{\sigma^2}{2}}] \tag{2}$$

In our experiment, $\sigma = \pi$. Each ψ_i is a plane wave characterized by the vector $\mathbf{k}_i$. The multiplicative factor $|\mathbf{k}_i|^2$ ensures that filters tuned to different frequency bands have approximately equal energies. The term $e^{-\frac{\sigma^2}{2}}$ is subtracted to render the filters insensitive to the overall level of lighting conditions . $\mathbf{k}_i$ is defined as

$$\mathbf{k}_i = |\mathbf{k}_i|(cos\varphi, sin\varphi) \tag{3}$$

$\mathbf{k}_i$ is parameterized by $|\mathbf{k}_i|$ and φ. The former decides the frequency of the kernel and the latter decides the orientation. We use six frequencies $\frac{\pi}{2}, \frac{\pi}{4}, \frac{\pi}{8}, \frac{\pi}{16}, \frac{\pi}{32} \frac{\pi}{64}$ and six orientations $\frac{\pi}{6}, \frac{\pi}{3}, \frac{\pi}{2}, \frac{2\pi}{3}, \frac{5\pi}{6}, \pi$. Because G_i is a complex value, only its amplitude $|G_i|$ is used to form the feature vector. This amplitude information degrades gracefully with shifts in the image location at which it is sampled. By concatenating, feature vectors from upper and lower faces form a 4968 dimension vector for later process.

3. Linear Discriminant Function

3.1. Principle Component Analysis

Principle component analysis, also known as Karhunen-Loeve methods, is often used in pattern recognition fields to reduce data dimensionality. Since it was used by Pentland [6] in early 90's for face recognition, it has been widely studied by face recognition research group. One of major advantages is that it requires no knowledge of class membership of pattern samples .It project the original feature vectors to a lower dimension along the directions where data have the largest variance, thus most information is retained.

Formally, given the d-D vectors $\{\mathbf{x_i}\}$, where $i = 1...n, d >> n$, the mean $\mathbf{m}$ and covariance $\mathbf{C}$ can be estimated as:

$$\mathbf{m} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x_i} \tag{4}$$

$$\mathbf{C} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x_i} - \mathbf{m})(\mathbf{x_i} - \mathbf{m})^\mathbf{t} \tag{5}$$

PCA tries to find the linear projection $\mathbf{W} = [\mathbf{w_1}...\mathbf{w_m}]$ such that the determinant of the covariance of the new projected vectors $\mathbf{y_i} = \mathbf{W^t x_i}$ is maximized. In other words, total scatter of the projected data is maximized, i.e.,

$$\mathbf{W} = argmax_\mathbf{w}|\mathbf{W^t CW}| \tag{6}$$

The problem is solved by noting that $\{\mathbf{w_i}|i = 1, 2...m\}$ is the set of d-D eigenvectors of $\mathbf{C}$ corresponding to the m largest eigenvalues.

3.2. Fisher Linear Discriminant

For the training set, the class label of each sample is known.So it might make sense to utilize this information. Given a set of d dimensional data $\{\mathbf{x}_i\}, i = 1, 2, ...n$ where n_i data in class $\omega_i, i = 1, 2, ...c$, Fisher linear discriminant tries to find a set of $c-1$ projections so that after projection, the ratio of the between class scatter over the total within class scatter is maximized. In detail, the total within class scatter $\mathbf{S_w}$ is defined as

$$\mathbf{S_i} = \sum_{\mathbf{x}\in\omega_\mathbf{i}}(\mathbf{x_i} - \mathbf{m_i})(\mathbf{x_i} - \mathbf{m_i})^\mathbf{t} \tag{7}$$

$$\mathbf{S_w} = \sum_{i=1}^{c}\mathbf{S_i} \tag{8}$$

where $\mathbf{m_i}$ is the mean of class ω_i. The between class scatter $\mathbf{S_b}$ is

$$\mathbf{S_b} = \sum_{\mathbf{i=1}}^{\mathbf{c}} \mathbf{n_i}(\mathbf{m_i} - \mathbf{m})(\mathbf{m_i} - \mathbf{m})^{\mathbf{t}} \quad (9)$$

where $\mathbf{m}$ is the total mean. FLD tries to find the linear projection $\mathbf{W} = [\mathbf{w_1}...\mathbf{w_{c-1}}]$ such that the ratio of the between class scatter over the total within class scatter is maximized of the new projected vectors $\mathbf{y_i} = \mathbf{W^t x_i}$ is maximized, i.e.,

$$\mathbf{W} = argmax_{\mathbf{w}} \frac{|\mathbf{W^t S_b W}|}{|\mathbf{W^t S_w W}|} \quad (10)$$

It can be shown that the above criterion is maximized by choosing $\mathbf{w_i}$ according to $\mathbf{S_w^{-1} S_b w_i} = \lambda_\mathbf{i} \mathbf{w_i}$. This is also an eigenvalue problem. However, in face emotion classification, the rank of $\mathbf{S_w}$ is at most $n-c$, which means $\mathbf{S_w}$ is singular. Similar to [7] we use PCA to project data to a lower dimension before apply FLD. If $\mathbf{W_{PCA}}$ denotes the transform matrix for PCA, $\mathbf{W_{FLD}}$ denotes the transform matrix for FLD, then the new criterion to be maximized is

$$\begin{aligned} J(\mathbf{W}) &= \frac{|\mathbf{W^t S_b W}|}{|\mathbf{W^t S_w W}|} \\ &= \frac{|\mathbf{W_{FLD}^t W_{PCA}^t S_b W_{FLD} W_{PCA}}|}{|\mathbf{W_{FLD}^t W_{PCA}^t S_w W_{FLD} W_{PCA}}|} \end{aligned} \quad (11)$$

The number of eigenvectors used in PCA can be evaluated empirically.

4. Neural Network

Many types of neural networks have been applied in facial emotion classification, including hop field network, radial basis function network and so on. Perhaps the most commonly used is multi-layer perceptron, a kind of neural network which can be thought of as a non-linear technique widely used in pattern classification, because of its discriminant power.

4.1. Single MLP

A single-hidden layer neural network is shown in Fig.2. It is fully connected between adjacent layers. The operation of this network can be regarded as a non-linear decision making process. Given an I dimensional input vector $\mathbf{x} = [\mathbf{x_1}, \mathbf{x_2}, ..., \mathbf{x_I}]'$ from one of class set $\omega_1, \omega_2, ..., \omega_N$, each output node yields the output value $y_o, o = 1, 2, ...N$ to that class by

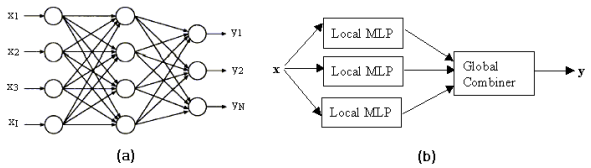

Figure 2: (a) single hidden layer MLP (b) multiple MLP classifier

$$y_o(W, \mathbf{x}) = f_o[\sum_{h=1}^{H} w_{ho} f_h(\sum_{i=1}^{I} w_{ih} x_i)] \quad (12)$$

where W is the whole set of all parameters in MLP, w_{ih} is the weight between input node i to hidden node h, w_{ho} is the weight between hidden node h to output node o, f_h and f_o are the transfer function used at hidden layer and output layer, respectively. Often non-linear function such as log sigmoid function and hyperbolic tangent function is used. The label of the input is evaluated by selecting the node having the largest value. The output of MLP as shown above is not just likelihood or binary logical values between zero and one. In fact, after reaching the global minimum on the error surface based on a certain cost function, it is the estimate of Bayesian posterior probability [1]. With a squared error cost function, the MLP is trained by adjusting its weights to minimize

$$E[\sum_{o=1}^{N} (y_o(\mathbf{x}) - d_o(\mathbf{x}))^2] \quad (13)$$

where $E[.]$ is the expectation operator, $y_o(\mathbf{x})$ and $d_o(\mathbf{x})$ are the actual and desired output values at output node o for input $\mathbf{x}$, respectively. Some manipulation of the above expression gives a form commonly used in statistics that provides insight for $y_o(\mathbf{x})$:

$$E[\sum_{o=1}^{N} (y_o(\mathbf{x}) - E[d_o|\mathbf{x}])^2] + E[\sum_{o=1}^{N} var[d_o|\mathbf{x}]] \quad (14)$$

where $E[d_o|\mathbf{x}]$ is the conditional expectation of d_o and $var[d_o|\mathbf{x}]$ is the conditional variance of d_o. For N-class problem, d_o is one if $\mathbf{x} \in \omega_\mathbf{o}$ and d_o is zero elsewhere. Hence, the conditional expectations are:

$$\begin{aligned} E[d_o|\mathbf{x}] &= \sum_{o=1}^{N} d_o P(\omega_o|\mathbf{x}) \\ &= P(\omega_o|\mathbf{x}) \end{aligned} \quad (15)$$

which is the Bayesian posterior probability. Therefore, for N-class problem, when MLP are trained to minimize the squared error cost function, the outputs can be regarded as the Bayesian probability so as to minimize the mean squared estimation error.

4.2. Multiple MLP Classifier

The MLP is trained on a set of samples and tries to find the optimal non-linear mapping to distinguish them. A single MLP of a finite size does not usually map the complete set, or over fitting on the training set while generalizes much worse on the test set. Seldom does it improve by simply increasing the size of hidden layer or number of hidden layers. What's more, in real application like facial emotion classification, the dimension of data is very large when employing Gabor wavelets as feature extraction method. The features are very complex, neither statistically independent nor unimodally distributed. So if we utilize multiple MLPs and let each of them only deals part of features or part of input space, the whole performance may be better. Using multiple small NN instead of one big NN is also supported in [2, 3]. In [2], an architecture, called hierarchical mixer of expert, was proposed, where each expert tried to catch the mapping for a specific region of input space. In [3], the strategy of one class at one network outperformed that of all classes at one network, both in result and computation cost.

The fundamental of multiple MLP is to develop n independent trained MLPs with specific features (see Fig.2). A global decider takes in the results from each of those MLPs and makes a decision with some fusion technique such as voting. However, too simple technique for combining is often not enough, so more sophisticated technique such as fuzzy logic or again, neural network can be considered.

5. Experimental Results

The database we work on consists of 100 facial image from 5 subjects, 3 females and 2 males. 50 images are for training and the remaining 50 are for test.

5.1. Linear Methods

We first normalize the original data set to make them zero mean. After obtaining the projection matrix $\mathbf{W}$ and getting the new vector $\mathbf{y} = \mathbf{W^t x}$, the mean $\mathbf{m_i}$ is computed for each class. Then, for classification, new pattern $\mathbf{x}$ is compared with the mean $\mathbf{m_i}$ based on the two criteria, Euclidean distance between two vectors and *cos* of the angle of two vectors:

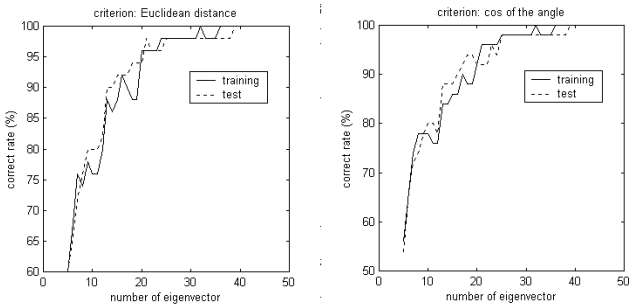

Figure 3: classification result with PCA

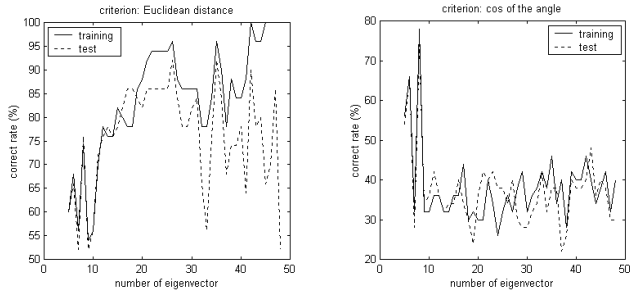

Figure 4: classification result with FLD

$$d_{dis} = |\mathbf{W^t}(\mathbf{x} - \mathbf{m}) - \mathbf{m_i}| \quad (16)$$

$$d_{cos} = \frac{\mathbf{x^t m_i}}{|\mathbf{x}||\mathbf{m_i}|} \quad (17)$$

Figures 3 show the classification rate with different number of eigenvector used.

It is shown as the number of eigenvector used in classification increases, the performance on both training and test set also improves.Correct rate reaches 100% from 39 eigenvectors.

Although FLD yields better results than PCA in face recognition [7], it does not outperform FLD in our experiment. With Euclidean distance criterion (see Fig.4), the results fluctuate considerably as more eigenvectors get used. Generally speaking, the correct rate reaches 100% from 45 eigenvectors, but the best result on test set is at the time when 35 eigenvectors get used and correct rates on training set and test set are 96% and 92% respectively. As for *cos* criterion, the performance deteriorates as more eigenvectors get used and the best result on test set is 74% when 8 eigenvectors used.

5.2. MLP

With single MLP, we set hidden layer size to 20 and each output node for one emotion. Log sigmoid transfer function is used at both hidden and output layer.

Before training, the whole data are normalized to have zero mean and unity standard deviation. The best result is 100% for training set and 96% for test set while it is much worse using unnormalized data.

With multiple MLP classifier, we divide the input vector, which consists of Gabor coefficients from a set of frequencies and orientations, into 6 parts two times, once according to frequency and the second time according to orientation. For each part of input vector at a specific frequency or orientation, a MLP is developed and trained separately to classify into 5 emotions. Their results are listed in Table 1, 2. At the global combiner level, we again employ a single layer perceptron, which receives the classification results from all 12 local MLPs and make a global decision. It achieves correct rate of 100% on both train and test set.

Table 1. **classification rate with different frequency**

frequency	$\frac{1}{2}\pi$	$\frac{1}{4}\pi$	$\frac{1}{8}\pi$	$\frac{1}{16}\pi$	$\frac{1}{32}\pi$	$\frac{1}{64}\pi$
training (%)	99	100	97	97	96	97
test (%)	96	95	93	93	92	93

Table 2. **classification rate with different orientation**

orientation	$\frac{1}{6}\pi$	$\frac{1}{3}\pi$	$\frac{1}{2}\pi$	$\frac{2}{3}\pi$	$\frac{5}{6}\pi$	π
training (%)	99	100	100	100	99	96
test (%)	93	94	96	97	90	92

6. Conclusion

This paper presents a comparative study of two types of techniques for facial emotion classification on the single image. In the first type of LDF, PCA and FLD are investigated. With PCA, very high classification rate is achieved when most eigenvectors are used. However, unlike the case in face recognition, the results become worse after we project the data obtained with PCA to a lower dimension with FLD, utilizing the class information of training samples. In the second type, we employ MLP, a non-linear technique and compare the performance by single big MLP and multiple MLP classifier. The experiment shows the latter gives better results than the former, without great effort to fine tune the individual MLP classifier. Although PCA and multiple MLP classifier give the best results, each technique has its own merit and gives some possibility to enlarge the conventional methods for facial emotion classification.

In addition, the experimental results with individual MLP in multiple MLP classifier show Gabor wavelets with high frequency may have slightly more facial information. This agrees with the work in [4]. It is also shown that Gabor wavelets having vertical orientations may be more important for facial emotion. A intuitive explanation for this is that most of major noticeable motions on face are vertical, such as eyebrow raiser and mouth opener.

References

[1] N. Morgan, H.A. Bourlard, "Neural Networks for Statistical Recognition of Continuous Speech", *Proc. IEEE* , Vol. 83, No. 5, pp. 742 -772, May 1995.

[2] R.A. Jacobs, M.I. Jordan, S.J. Nowlan, G.E. Hinton, "Adaptive Mixtures of Local Experts", *Neural Computation* , Vol. 3, pp. 79-87, 1991.

[3] S.Y. Kung, M. Fang, S.P. Liou, M.Y. Chiu, J.S. Taur, "Decision-Based Neural Network for Face Recognition System", *Proc. International Conference on Image Processing* , Vol. 1, pp. 430 -433, 1995.

[4] G. Donato, M.S. Bartlett, J.C. Hager, P. Ekman, T. J. Sejnowski, "Classifying Facial Actions", *IEEE Trans. Pattern Analysis and Machine Intelligence* , Vol. 21, No. 10, pp. 974 -989, Oct. 1999.

[5] Z. Zhang, M. Lyong, M. Schuster, S. Akamatsu, "Comparison Between Geometry-Based and Gabor-Wavelets-Based Facial Expression Recognition Using Multi-Layer Perceptron", *Proc. Third IEEE International Conference on Automatic Face and Gesture Recognition* , pp. 454 -459, 1998.

[6] M. Turk, A. Pentland, "Eigenfaces for Recognition", *Proc. IEEE Conference on Computer Vision and Pattern Recognition* , pp. 586-591, 1991.

[7] P.N. Belhumeur, J.P.Hespanha, D.J. Kriegman, "Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection", *IEEE Trans. Pattern Analysis and Machine Intelligence* , Vol. 19, No. 7, pp. 711 -720, Jul. 1997.