13-6

MVA2007 IAPR Conference on Machine Vision Applications, May 16-18, 2007, Tokyo, JAPAN

Facial Expression Recognition by SVM-based Two-stage Classifier on Gabor Features

Fan CHEN, Kazunori KOTANI
School of Information Science
Japan Advanced Institute of Science and Technology
Ashahi-dai 1-8, Nomi, Ishikawa 923-1292, Japan
chen-fan@jaist.ac.jp ikko@jaist.ac.jp

Abstract

We propose a two-stage classifier for the elastic bunch graph matching based recognition of facial expressions. The major purpose is to calculate distinctive similarity between image patterns by applying optimal weights to responses from different Gabor kernels and those from different fiducial points. In the first stage, we perform SVM on each fiducial point individually to extract a weighted feature from the Gabor response. The optimal fusion of those features is then calculated by another stage of SVM, providing the weight between fiducial points. From numerical experiments, the proposed method shows improved performances when comparing with other methods.

1 Introduction

Normalization error in head pose variations and fiducial point displacements significantly affects the performance of appearance-based methods for both face and facial expression recognition, e.g., eigenface, Fisherface, etc. Some model-based methods such as the Elastic Bunch Graph Matching(EBGM)[1], could alleviate the problem by explicitly including the position matching of features. Using Gabor filters in EBGM to extract local descriptors further improves the robustness against light and contrast.[2] [3] Extensions of EBGM to tolerate larger variations in pose and to handle larger dataset have also been proposed.[4] After the extraction of local descriptors, the similarity between two images is computed for facial expression classification. We observe that not only some fiducial points but also some responses from certain Gabor filters are more distinctive than others in discriminating different facial expressions. Instead of using the simple sum of local similarity values at each fiducial point, it is reasonable for us to apply weighted sum to both responses from different Gabor kernels and local descriptors at different points. Many statistical methods have been applied to the EBGM method for searching the optimal weights between fiducial points, which include Linear Discriminant Analysis(LDA)[5][6], Support Vector Machine(SVM)[7][8] and neural networks[11]. From the viewpoint of combining classifiers, they can be regarded as performing the output fusion on the bank of Gabor filters.[9][10] If we extract a further distinctive vector from the Gabor jet at each point, further improvements on accuracy are available. Direct classification of the feature vector constructed from all local descriptors of each image might cause over-fitting, due to the curse of dimensionality. We consider a two-stage classifier for facial expression recognition, which applies base classifiers on all fiducial points individually to estimate the optimal weight of response from different Gabor kernels and then the fusion of their outputs is calculated by a second-stage classifier to finally obtain the classification support. Especially, we use SVM in both the base classifiers and the second-stage classifier for its good generalization ability.

In Section 2, we introduce briefly the architecture of our two-stage classifier. In Section 3, numerical experiments are made to investigate the performance of the proposed method. Comparisons with other methods are also made. Concluding remarks and comments on our future works are given in Section 4.

2 SVM-based Two-stage Classifier for Facial Expression Recognition

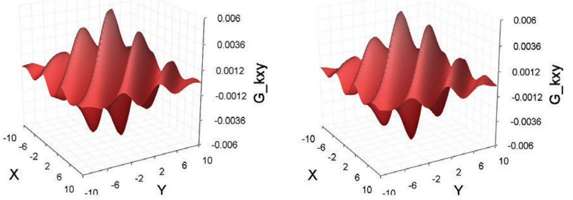
Our major purpose is to find a optimal classifier on local descriptors gathered from EBGM. Therefore, we will omit the introduction of EBGM and focus the explanation on the development of classifier on obtained local descriptors. We let $\mathbf{q}^{(n)}$ be the n-th image vector in a set of N images $\mathbf{Q} = \{\mathbf{q}^{(n)}|n \in \{1, \cdots, N\}\}$. $\mathbf{z} = \{z_n|n \in \{1, \cdots, N\}\}$ is the set of class labels with $z_n \in \{1, \cdots, C\}$ being the label of the n-th image vector. C is the number of all classes. We extract M fiducial points for the n-th image manually or automatically, whose coordinates form the set $\{(x_m^{(n)}, y_m^{(n)})|m \in \{1, \cdots, M\}\}$. Local descriptors at fiducial points are extracted by a Gabor bank of K filters, $\mathbf{G} = \{\mathbf{G}_k|k \in \{1, \cdots, K\}\}$ in Ref.[3], where the element at coordinate (x, y) of the k-th filter reads,

$$G_{kxy} = \frac{\exp\left[-\frac{1}{2}\left(\frac{x^2}{\sigma_{kx}^2}+\frac{y^2}{\sigma_{ky}^2}\right)\right]}{2\pi\sigma_{kx}\sigma_{ky}}\left\{\exp\left[i\left(\omega_{kx}x+\omega_{ky}y\right)\right]\right. \\ \left. -\exp\left[-\frac{\left(\omega_{kx}\sigma_{kx}\right)^2+\left(\omega_{ky}\sigma_{ky}\right)^2}{2}\right]\right\}, \quad (1)$$

which is a complex sinusoid centered at frequency $(\omega_{kx}, \omega_{ky})$ and modulated by a Gaussian envelope, as shown in Fig.1. σ_{kx} and σ_{ky} are the standard deviations of the elliptical Gaussian along x and y. The second term inside the bracket makes the Gabor kernel DC-free to improve the robustness against brightness variation. A Gabor jet is defined as the vector of Gabor response for the m-th fiducial point, whose k-th element is the convoluted result between the image and the k-th Gabor kernel under offset $(x_m^{(n)}, y_m^{(n)})$, i.e.,

$$u_{mk}^{(n)} = \sum_x \sum_y \mathbf{q}^{(n)}_{x_m^{(n)}-x, y_m^{(n)}-y} G_{kxy}. \tag{2}$$

$\mathbf{q}^{(n)}_{x,y}$ is the intensity at pixel (x, y) of the n-th image.



(a.) Real part (b.) Imaginary part

Fig. 1: A Gabor kernel.

Since the feature vector constructed from all local descriptors of each image is in general of a longer length than the size of the training set, direct classification of this feature vector might cause over-fitting, due to the curse of dimensionality. We consider a two-stage classifier, whose diagram is given in Fig. 2. The weight of responses from different Gabor kernels are optimized by first-stage classifiers, the fusion of whose outputs is computed by the second-stage classifier to produce the classification support.

1.) For the n-th image, EBGM is used to search optimal coordinates $\{(x_m^{(n)}, y_m^{(n)})\}$ for all fiducial points by performing local searches inside a neighboring area of a starting point next to the ground truth. Detailed explanations on EBGM can be found in Ref.[1] and [3].

2.) For the located m-th fiducial point, its Gabor jet is calculated and normalized to unit length, i.e., $\mathbf{u}_m^{(n)} = [|u_{mk}^{(n)}|/\sqrt{\sum_k |u_{mk}^{(n)}|^2} | k \in \{1, \cdots, K\}]^T$. The magnitude normalization of Gabor jets improves the robustness against light and contrast variations.

3.) Gabor jets from different training data at the m-th fiducial point form a Gabor bunch $\{\mathbf{u}_m^{(n)} | n \in \{1, \cdots, N\}\}$ to train the m-th max-win SVM classifier at the first stage. In the testing phase, Gabor Jet $\mathbf{u}_m^{(n)}$ is fed into the trained max-win SVM classifier which outputs $\mathbf{v}_m^{(n)} = [v_{md}^{(n)} | d \in \{1, \cdots, D\}]^T$ with D being the output dimensionality.

4.) Outputs for the n-th image from all M classifiers can be reorganized in a decision profile $\mathbf{V}^{(n)} = [\mathbf{v}_m^{(n)} | m \in \{1, \cdots, M\}]^T$, which is vectorized and sent to the pairwise SVM classifier for classification.

Different kinds of combiners on the fusion of decision profile have been proposed, which can be classified as non-trainable and trainable combiners[9].

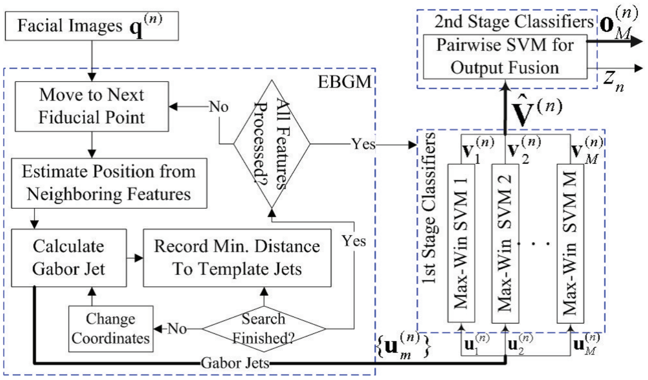


Fig. 2: Block diagram for proposed two-stage classifier.

Average-fusion and max-fusion, two commonly used non-trainable combiners, are defined as:

$$\text{average}-\text{fusion} \quad : \quad o_d^{(n)} = \frac{1}{M}\sum_m v_{md}^{(n)}, \hat{D} = D, \tag{3}$$

$$\max-\text{fusion} \quad : \quad o_d^{(n)} = \max{}_m v_{md}^{(n)}, \hat{D} = D, \tag{4}$$

respectively. $\mathbf{o}^{(n)} = \{o_{\hat{d}}^{(n)} | \hat{d} \in \{1, \cdots, \hat{D}\}\}$ is the classification support from the combiner. Different from the usual weighted average fusion,

$$\mathbf{o}^{(n)} = \frac{1}{M}\sum_m w_m v_{md}^{(n)}, \hat{D} = D, \tag{5}$$

we consider a generalized weighted average fusion, i.e.,

$$\mathbf{o}^{(n)} = \mathbf{w}\hat{\mathbf{V}}^{(n)}, \tag{6}$$

where $\hat{\mathbf{V}}^{(n)} = [[\mathbf{v}_m^{(n)}]^T | m \in \{1, \cdots, M\}]^T$ is the vectorized form of $\mathbf{V}$.

Various statistical classifiers are applicable to the second-stage classifier to calculates the optimal fusion matrix $\mathbf{w}$ of the decision profile. In both stages of classifiers, we take SVM whose purpose in a two-class problem is to maximize the margin between two classes, or equaivalently, to search for $\mathbf{w}$ that minimizes the following objective function:

$$L_p = \{\frac{1}{2}\mathbf{w}^T\mathbf{w} - \sum_n \alpha_n [z_n^*(\mathbf{w}^T\hat{\mathbf{V}}_m^{(n)} + w_0) - 1]\}, \tag{7}$$

where $\{\alpha_n | n = 1, \cdots, N, \alpha_n > 0\}$ are the Langrange multipliers and z_n^* takes 1 for $z_n = 0$ and -1 for $z_n = 1$. Detailed explanations on SVM can be found in Ref.[7].

When extending the two-class SVM to the multi-class version, a Max-Win multi-class SVM adopts the one-against- rest strategy, while a Pairwise one takes the one-against-one approach. We use the Max-Win multi-class SVM for the first stage classifiers because it produces continuous output. Pairwise multi-class SVM has been used in the second stage of our method because it provides better performance, based on numerical data which is not shown in the present paper.

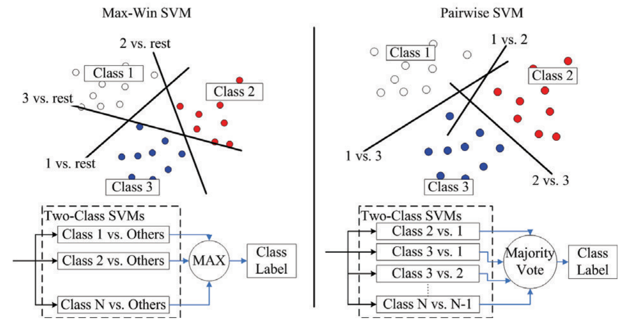


Fig. 3: Two kinds of multi-class SVM: Max-Win SVM and Pairwise SVM.

3 Numerical Experiments and Discussions

We focus on the performance comparisions between the proposed method and other fusion methods. The single stage versions of classifiers will also be compared. The Japanese Female Facial Expression (JAFFE) Database [2] is adopted, which includes 213 images in total. The goal of recognition is to classify them into neutral face or one of six elemental facial expressions suggested by Ekman et al.[12], i.e., happiness, anger, fear, disgust, sadness and surprise. We normalize these images through the alignment of both eye positions for later comparisons with appearance-based SVM. Some normalized samples are given in Fig. 4. All images are resized to 100 × 120 pixels.

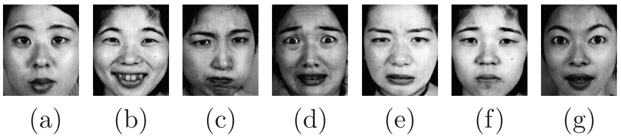


Fig. 4: Some normalized samples that are used in our numerical experiments from the JAFFE database. (a) Neutral (b) Happiness (c) Anger (d) Fear (e) Disgust (f) Sadness and (g) Surprise.

Fiducial points are manually landmarked in our experiments for both the training and testing dataset. 52 selected fiducial points is given in Fig.5, with 30 first-tier points (marked in crossing symbol) and 22 second-tier points (marked in circles). First-tier points are independently located while second-tier points are calculated from the positions of its neighboring first-tier points by taking centering or crossing points. We refer to the case of using LDA as the second-stage classifier as LDA-fusion, while we name the case of using SVM as SVM-fusion. Further, methods of applying SVM directly to the feature vectors from raster-scanned pixels or from the Gabor jets at all fiducial points of each image are referred to as appearance-based SVM method, stacked Gabor jet based SVM method, respectively.

To evaluate the performance quantitatively, we define a recognition rate as

$$r_c = \frac{1}{N}\sum_{n=1}^{N}\delta(z_n, z_n^*) \qquad (8)$$

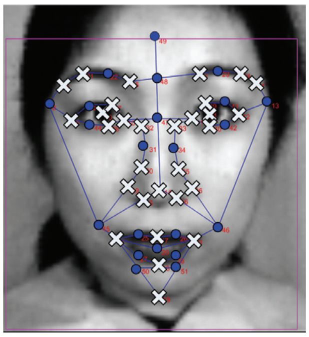


Fig. 5: Location of selected fiducial points. Crossing symbols are for first-tier fiducial points and circles for second-tier fiducial points.

where $\delta(x, y)$ is the Kronecker delta. z_n^* is the estimated label value. Due to the limited size of training dataset, the recognition rate on the training data is 1.0 or near 1.0 for almost all cases. Therefore, we will only give the results on testing dataset and mainly focus on the comparison of their generalization ability to untrained testing data. Numerical experiments have been performed on 15 randomly selected training sets with size N varying from 44 to 100. For each N, a pair of two mutually exclusive sets was created, one with N images for training, and another with $213 - N$ images for testing. In total, 15 pairs of training and testing datasets are used in our experiments and their results are summarized as follows:

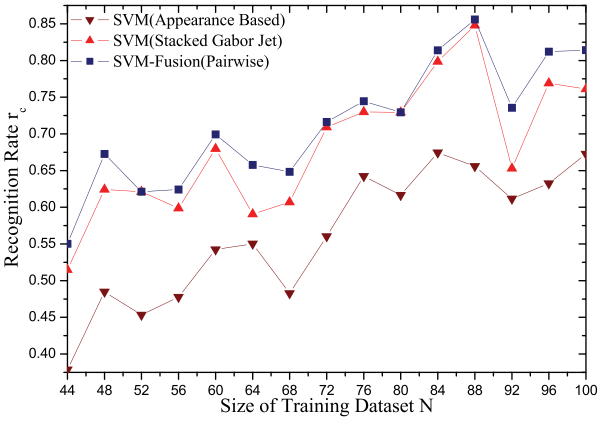


Fig. 6: Recognition rate r_c on the testing dataset for the appearance based SVM method, the stacked Gabor jet based SVM method and the SVM-fusion method. The proposed SVM-fusion achieves the best performance in recognition.

1.) In Fig.6, recognition rate r_c is plotted as a function of N, the size of training set, for three methods, i.e., the appearance-based SVM method, the stacked Gabor jet based SVM method and the SVM-fusion method. Comparing with the appearance-based SVM, the proposed method significantly increased the accuracy of recognition by including explicit position-matching of fiducial points. We also note that r_c from the two-stage SVM is about 3.1 points higher than that in the stacked Gabor jet-based SVM method on average.

2.) Recognition rates r_c under different fusion methods are compared in Fig.7. Average-fusion(AVE-Fusion), Max-fusion, LDA-fusion, and the proposed SVM-fusion have been tested. We found that SVM-fusion has the highest accuracy of recognition, which is considered as a result of the better generalization ability of SVM when comparing to LDA.

3.) In above experiments, we use linear kernels in all SVM implementations. In Fig.8, a comparison is made between SVM-fusion with a linear kernel and a Gaussian RBF kernel, i.e., $\phi(x_i, x_j) = \exp(-\gamma|x_i - x_j|^2)$. We take $\gamma = 0.1$ for this figure. For small training datasets nonlinear version may cause overfitting which damages its ability of generalization, while with enough training samples the nonlinear SVM-fusion outperforms that using a linear kernel.

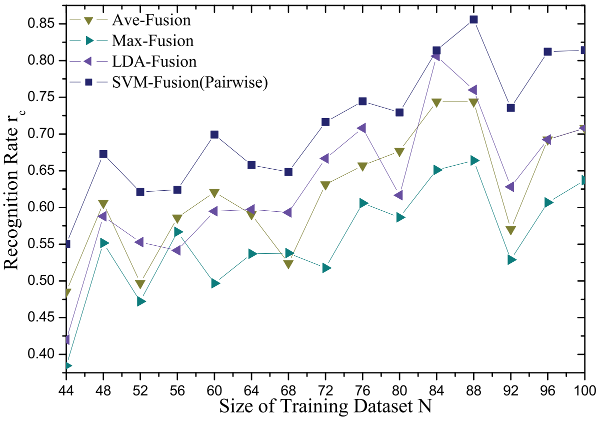


Fig. 7: Recognition rate r_c on the testing dataset under different fusion methods.

Although only one sample set for each size of bunch has been tested, we still conclude that the proposed method has better performance, because the proposed method outperforms other methods for all sizes of testing datasets we have tested.

4 Conclusions

We have proposed a method to enhance fiducial-point based recognition of facial expressions by estimating optimal weights on both different Gabor kernels and different fiducial points. We achieved this goal by developing a two-stage classifier. Max-win multi-class SVMs have been used as the base classifiers for producing continuous output and a pair-wise multi-class SVM has been applied for output fusion. Numerical experiments on manually landmarked fiducial points verified the efficiency of our proposed method in facial expression recognition. Further numerical experiments on the robustness against the localization error of EBGM will be made in our future works. We will also consider the possibility of including bagging for base classifier selection and utilizing the decision tree to organize base classifiers for further improvements.

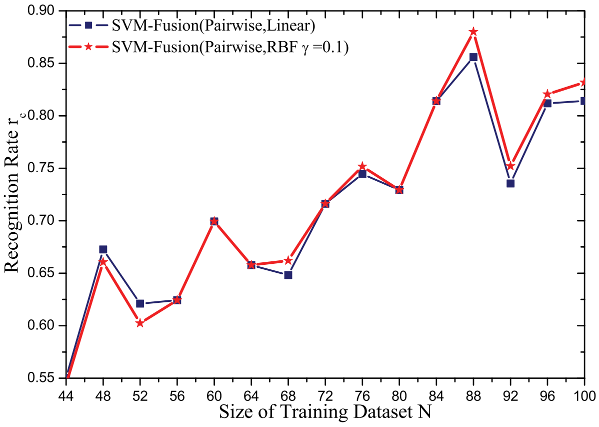


Fig. 8: Recognition rate r_c on the testing dataset under linear and nonlinear SVM-fusions. The nonlinear SVM uses a RBF kernel with $\gamma = 0.1$.

References

[1] J. Zhang, Y. Yan, and M. Lades, "Face recognition: eigenface, elastic matching and neural nets," *Proceedings of the IEEE,* vol.85, pp.1423-1435, 1997.

[2] M.J. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba, "Coding facial expressions with Gabor wavelets," *Proc. 3rd IEEE Int'l Conf. on Automatic Face and Gesture Recognition,* vol.1, pp.200-205, 1998.

[3] L. Wiskott, J.M. Fellous, N. Kruger, and C. Malsburg, "Face recognition by elastic bunch graph matching," *IEEE Trans. PAMI,* vol.19, pp.775-779, 1997.

[4] R.P. Wurtz, "Object recognition robust under translations, deformations and changes in background," *IEEE Trans. PAMI,* vol.19, pp.769-775, 1997.

[5] C. Liu, and H. Wechsler, "Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition," *IEEE Trans. Image Processing,* vol.11, pp.467-476, 2002.

[6] Y. Pang, L. Zhang, M. Li, Z. Liu, and W. Ma, "A novel Gabor-LDA based face recognition method," *Lecture Notes in Computer Science,* vol.3331, pp.352-358, 2004.

[7] C.J.C. Burges. "A tutorial on support vector machines for pattern recognition," *Data Mining and Knowledge. Discovery,* vol.2 no.2 pp.955-974, 1998.

[8] A. Tefas, C. Kotropoulos, and I. Pitas, "Using support vector machines to enhance the performance of elastic graph matching for frontal face authentication," *IEEE Trans. PAMI,* vol.23, pp.735-746, 2001.

[9] L.I. Kuncheva, "Combining pattern classifiers: methods and algorithms," Wiley, 2004.

[10] W.F. Liu, and Z.F. Wang, "Facial expression recognition based on fusion of multiple Gabor features," *Proc. 18th IEEE Int'l Conf. on Pattern Recognition (ICPR'06),* vol.1, pp.536-539, 2006.

[11] N. Ueda, "Optimal linear combination of neural networks for improving classification performance," *IEEE Trans. PAMI,* vol.22, pp.207-215, 2000.

[12] P. Ekman, W.V. Friesen, and P. Ellsworth, "Emotion in the human face: Guidelines for research and an integration of findings," *Pergamon Press,* New York, 1972.