ELectronics EXpress

IEICE Electronics Express, Vol.5, No.24, 1061–1066

Improving predictive accuracy by evolving feature selection for face recognition

Nan Liu and Han Wang[a)]

Electrical and Electronic Engineering, Nanyang Technological University, 50 Nanyang Avenue, Singapore 639798

a) *hw@ntu.edu.sg*

Abstract: Face recognition system usually consists of feature extraction and pattern classification. However, not all of extracted facial features contribute to the classification positively because of the variations of illumination and poses in face images. In this paper, an evolutionary feature selection algorithm is proposed in which discrete cosine transform (DCT) and genetic algorithms (GAs) are utilized to create a framework of feature acquisition. In detail, the face images are first transformed to frequency domain through DCT, then GAs are used to seek for optimal features in the redundant DCT coefficients where the generalization performance guides the searching process. Furthermore, an entropy-based extension on proposed evolving feature selection method is presented. In experiments, two face databases are used to evaluate the effectiveness of our proposals.

Keywords: feature selection, genetic algorithms, discrete cosine transform, entropy, face recognition

Classification: Science and engineering for electronics

References

[1] Z. H. Sun, G. Bebis, and R. Miller, "Object detection using feature subset selection," *Pattern Recognition*, vol. 37, pp. 2165–2176, 2004.

[2] M. J. Er, W. L. Chen, and S. Q. Wu, "High-speed face recognition based on discrete cosine transform and rbf neural networks," *IEEE Trans. Neural Netw.*, vol. 16, pp. 679–691, March 2005.

[3] F. Tan, X. Z. Fu, Y. Q. Zhang, and A. G. Bourgeois, "A genetic algorithm-based method for feature subset selection," *Soft Computing*, vol. 12, pp. 111–120, 2007.

[4] H. Liu and L. Yu, "Toward integrating feature selection algorithms for classification and clustering," *IEEE Trans. Knowl. Data Eng.*, vol. 17, pp. 491–502, 2005.

[5] C. J. Liu, "Evolutionary pursuit and its application to face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, pp. 570–582, 2000.

[6] M. Dash and H. Liu, "Handling large unsupervised data via dimensionality reduction," *Proc. ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery*, 1999.

DOI: 10.1587/elex.5.1061
Received October 17, 2008
Accepted November 17, 2008
Published December 25, 2008

IEICE Electronics Express, Vol.5, No.24, 1061–1066

1 Introduction

Features play a very important role in the task of pattern classification. Consequently, the selection of suitable features is necessary as most of raw data might be redundant or irrelevant to the recognition of patterns. In some cases, the classifier cannot perform well because of the large number of redundant features [1]. Face recognition is a challenge application of pattern recognition, and numerous efforts have been put on the extraction of discriminatory facial features. From geometry-based features to statistical features obtained from subspace analysis, one crucial step of face recognition is the facial feature extraction. In recent years, some researches have investigated the possibility of extracting features in frequency domain, such as the space obtained by performing discrete cosine transform (DCT) [2]. The results have shown that it is feasible and promising to extract discriminatory frequency components for classification. It is also worth paying attention that even the most dominant features may degrade the system performance because of the existence of variations on illumination or pose in face images. However, it is hard to determine which feature components are bounded with specific factors [4]. Naturally, selecting proper features is considered as one important step prior to classification. Given a set of d features, the task of selecting a subset of size m that leads to the smallest generalization error is defined as feature selection problem. A common procedure to discover optimal subset is examining all $\binom{d}{m}$ possible subsets and evaluating them with classification accuracy as criterion. Nevertheless, exhaustive searching method is impractical for even moderate values of m and d. Consequently, a number of algorithms are proposed to find near optimal subsets using non-exhaustive sequential feature selection procedure. Sun et al. [1] and Tan et al. [3] proposed using genetic algorithms (GAs) to select features for classification. In this paper, a GAs-based feature selection algorithm using novel genetic coding scheme is proposed and an extension based on entropy is also introduced to improve predictive accuracy. Details are described in the following sections.

2 Proposed evolving feature selection in frequency domain

In this section, a framework of evolutionary feature selection in frequency domain is proposed. At first, DCT is implemented to convert the image into frequency domain and the feature dimensionality is reduced by discarding high frequency components. Then, GAs are used to select discriminatory features from the DCT coefficients under the driven of fitness function. In addition, an entropy-based extension is presented.

2.1 Discrete cosine transform

DCT is a widely used feature extraction technique that converts images into frequency domain and extracts features by discarding high-frequency coefficients. The extracted features by DCT are useful for face recognition as significant facial features such as the outlines of hair and face, positions of eyes, nose, and mouse, are maintained [2]. Given an image $f(s,t)$ of size

DOI: 10.1587/elex.5.1061
Received October 17, 2008
Accepted November 17, 2008
Published December 25, 2008

$M \times N$, where $1 \le s \le M$ and $1 \le t \le N$, the DCT of image is defined by

$$F(u,v) = \alpha(u)\alpha(v) \sum_{s=0}^{M-1} \sum_{t=0}^{N-1} f(s,t) \cos\left[\frac{\pi(2s+1)u}{2M}\right] \cos\left[\frac{\pi(2t+1)v}{2N}\right] \quad (1)$$

where $\alpha(u)$ and $\alpha(v)$ are $\frac{1}{\sqrt{M}}$ and $\frac{1}{\sqrt{N}}$ if u and v are zeros; otherwise they are $\sqrt{2}$ times larger for non-zero variables. In application, DCT is applied on the entire image so as to avoid the risk of loosing relationship among components of image, and thus the transform shares the same dimension as original image. Given the fact that the DCT coefficients with large magnitude are located in the upper-left corner, we use an $n \times n$ window to extract coefficients to convert 2D DCT matrix into feature vector to represent original image.

2.2 Genetic algorithms-based feature selection

In the extracted features by DCT, dominant coefficients may not correlate to high discriminatory power, hence it is hard to model the feature space to discover the correlations among the DCT components and variations of poses and illumination in the images. Considering the difficulties mentioned above, genetic algorithms are proposed to explore the feature space that contains complex interacting components to select most fitting features in terms of classification accuracy. In brief, representations of candidates in the population are the concatenation of various binary strings. The proposed encoding scheme is shown in Fig. 1. In this example, 3×3 DCT coefficients are extracted, thus nine genetic codes $C_1, C_2, \ldots, C_9$ are needed for feature selection. In experiments, the codes are generated randomly within the bound of $[0, 1]$, and 22 bits string is used to represent each code, i.e., the nine binary strings are concatenated together to form a chromosome with the length of 198. In our proposal, a proper threshold ε is required to assign the codes to two classes: The codes in white blocks ($C_i \ge \varepsilon$) mean that the corresponding DCT coefficients are retained in feature vector; on the contrary, the coefficients associated with codes in gray background ($C_i < \varepsilon$) will be discarded. For the example in Fig. 1, the values of C_2, C_5, C_7 are smaller than ε, thus the

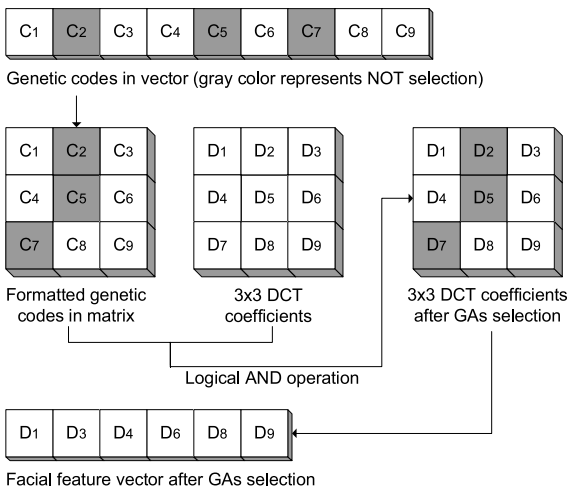

Fig. 1. Schema of genetic coding mechanism.

DOI: 10.1587/elex.5.1061
Received October 17, 2008
Accepted November 17, 2008
Published December 25, 2008

DCT coefficients D_2, D_5, D_7 are excluded from the feature vector. The white and gray blocks can be simply defined as logical "1" and "0". Alternatively, the genetic codes C_i can be encoded as "1" or "0" directly [1]. However, such coding scheme might fail in controlling feature dimensionality because it has the same probability to get "1" or "0". By introducing the threshold ε, the proportion of preserved DCT coefficients will be adjustable, that is, the probability of getting "1" can be increased or reduced as practical needs.

The fitness function guides the entire converging process. In most evolutionary pattern recognition systems, classification accuracy (CA) on training set is employed as the fitness [1, 5]. To avoid overtraining, the classification accuracy on K-fold cross-validation is used for measuring fitness, i.e., $\zeta(\mathcal{F}) = \frac{1}{K}\sum_{i=1}^{K} \mathrm{CA}_i$ where K is 10 in this paper. During the iterations of several genetic operations, a predefined number of generations is deployed to terminate the evolving procedure. The optimal genetic codes obtained in the last generation will be used to select features for testing.

2.3 Wrapper plus filter: an entropy-based extension

According to evaluation criterion, the feature selection methods can be separated into wrapper and filter models in terms of whether or not induction methods are enrolled [4]. In this section, an entropy-based filter model [6] is incorporated into the evolving feature selection process (wrapper). The fact that the dataset with distinct clusters has low entropy and data of chaotic configurations results in a large entropy value, promises that dataset includes less irrelevant features when a low entropy is observed. The entropy measure of a dataset of N instances $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$ is calculated as follows

$$\mathrm{E} = -\sum_{i=1}^{N}\sum_{j=1}^{N}(S_{ij} \times \log S_{ij} + (1 - S_{ij}) \times \log(1 - S_{ij})) \tag{2}$$

where $S_{ij} = e^{-\alpha \times D_{ij}}$ is the similarity measure based on distance between two samples $\mathbf{x}_i$ and $\mathbf{x}_j$ with all numeric features. $\alpha = \frac{-\ln 0.5}{\overline{D}}$ is a parameter and D_{ij} is the Euclidean distance between the two instances. $\overline{D}$ is the average distance among samples in data set. In our proposal, entropy values are calculated for all candidates in population. As large entropy indicates disorderly configurations in data, the corresponding fitness of feature vector is set to a small value (0.1) to reduce the effect of disorderly data on classification.

3 Experiments

Extraction and selection of facial features is emerging as an active area in the field of face recognition. Despite numerous papers put efforts on using complex and powerful classifiers, more research works are focusing on the acquiring of statistically dominant and discriminatory features. Consequently, a GAs-based feature selection scheme is proposed to pick up significant variables in frequency domain. In order to validate the method, experiments are set up with ORL face databases, and a combo database consisting of ORL, Yale, and UMIST datasets. In ORL database, 200 images of 40 individuals

DOI: 10.1587/elex.5.1061
Received October 17, 2008
Accepted November 17, 2008
Published December 25, 2008

are used for training, and the remaining 200 images form the testing set. The combo set contains 555 images for training and 575 face images for testing.

Prior to classification, the dimensionality of selected features are further reduced using the state-of-the-art subspace methods principal component analysis (PCA), linear discriminant analysis (LDA) and kernel PCA (KPCA). Subsequently, 1-nearest neighbor algorithm (1-NN) is used to predict labels for samples, where the nearest neighbors of an pattern are defined in terms of the Euclidean distance. In GAs, population size, crossover probability p_c, and mutation probability p_m are set to 50, 0.65, and 0.004, respectively. Moreover, 100 generations is considered as the termination criterion.

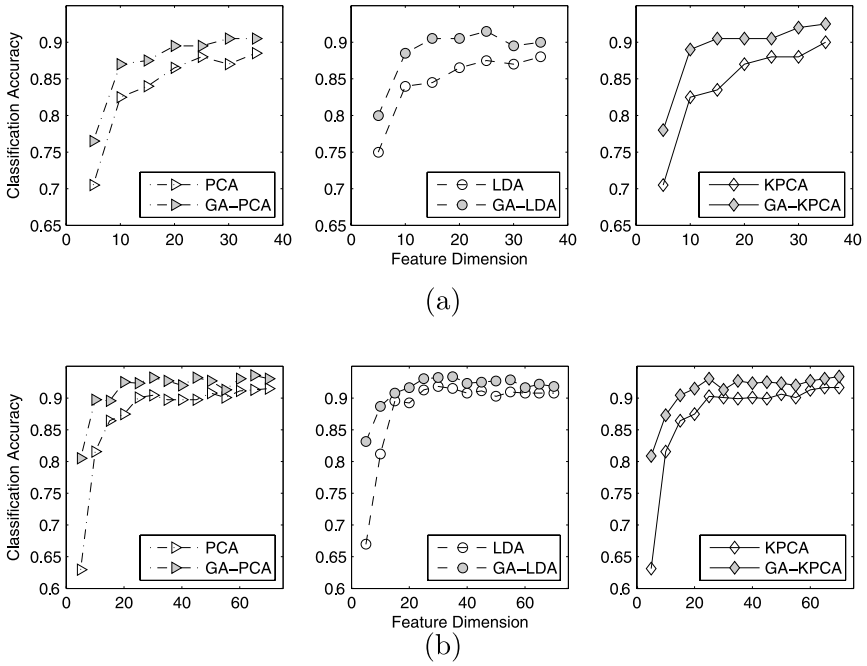

Fig. 2. (a) Experimental results on ORL face database; (b) experimental results on combo face database.

Fig. 2 (a) depicts the comparison results on ORL database. GA-PCA, GA-LDA, and GA-KPCA represent the extracted features with different dimensionality reduction methods that are implemented on selected features through evolving. It is obvious that the proposed facial features present high discriminatory power than those features without evolutionary selection. Fig. 2 (b) shows the results on combo face database, in which the features obtained by evolving selection by GAs achieve good performance across all dimensions. In frequency coefficients extraction, DCT reduces the image dimension to 81 (9×9 block) for ORL dataset and 121 (11×11 block) for combo database. In addition, the threshold ε is set to 0.3 to ensure that the dimension of frequency components after evolutionary selection is larger than the expected dimensionality of features. A small threshold can retain more dominant DCT coefficients and thus might achieve satisfied generalization performance. In practice, a trade-off between the algorithm efficiency

DOI: 10.1587/elex.5.1061
Received October 17, 2008
Accepted November 17, 2008
Published December 25, 2008

Table I. Results of entropy-based extension on ORL.

$\zeta(\mathcal{F})$	CA	CA′			
m		$\mu = 1$	$\mu = 2$	$\mu = 3$	$\mu = 4$
35	0.905	0.920	**0.940**	0.910	0.910
30	0.905	0.915	**0.920**	0.895	0.895
25	0.895	0.890	**0.900**	0.895	0.895
20	0.895	0.885	**0.900**	0.880	0.880
15	0.875	**0.885**	**0.885**	0.880	0.865
10	0.870	**0.885**	**0.885**	0.875	0.860

and classification accuracy should be considered. Consequently, small block size is adopted to decrease the burden of computation.

The results in Table I show the effects of entropy-based strategy, in which PCA-based features are implemented. The highest accuracy on each feature dimension is highlighted in bold font. CA and CA′ indicate different fitness calculation schemes, and μ indicates how many candidates associated with top entropy values in population are reset with small fitness values. It can be seen that entropy-based modifications perform dominantly in generalization. With the increasing of μ, accuracy decreases possibly because some evolved feature candidates that are able to achieve high recognition rates have fairly large entropy values.

4 Conclusions

This paper presents a novel evolving feature selection algorithm taking the advantages of discrete cosine transform and genetic algorithms to improve predictive accuracy. The first step of proposed method is to convert images to comparably low-dimensional frequency component vectors using DCT. GAs are subsequently used to search optimal combination of DCT coefficients. The architecture of obtaining discriminative features is completed by applying the state-of-the-art dimensionality reduction methods such as PCA, LDA, and KPCA. Experimental results reveal that the evolutionarily selected features perform better than original DCT coefficients on face databases. In addition, facial features acquired by the hybrid model-based feature selection (wrapper + filter) show their superiority on ORL face database as well.

Acknowledgments

The authors would like to thank the researchers who have kindly provided the ORL, Yale and UMIST face databases. The authors are also grateful to all the reviewers and editor for their valuable comments and suggestions.

DOI: 10.1587/elex.5.1061
Received October 17, 2008
Accepted November 17, 2008
Published December 25, 2008