A Coarse Classification Scheme Based On Clustering and Distance Thresholds

Te-Wei Chiang[1] Tienwei Tsai[2]

[1]Department of Accounting Information Systems, Chihlee Institute of Technology, Taipei, Taiwan, R.O.C.
[2]Department of Information Management, Chihlee Institute of Technology, Taipei, Taiwan, R.O.C.
[1]ctw@mail.chihlee.edu.tw [2]twt@mail.chihlee.edu.tw

Abstract

In this paper a cluster-based coarse classification scheme is proposed to speed up the classification process. In order to develop a coarse classification scheme which is low dependent on domain-specific knowledge, discrete cosine transform (DCT) is employed to extract general and reliable features for vision-oriented applications, such as optical character recognition (OCR). Our coarse classification scheme is based on *k*-means clustering. On classifying an unknown object, the classes within the clusters that satisfy some thresholds are chosen as candidates. In this way, a large number of improbable candidates can be eliminated at the earlier stage, and only few surviving candidates need to be further examined in the subsequent fine classification process. This scheme can significantly reduce the classification time while maintaining the classification accuracy. Some experimental results are given to show the validity of our approach in the case of recognizing handwritten characters in Chinese paleography.

Keywords: Coarse classification, clustering, discrete cosine transform, optical character recognition.

1. Introduction

Classification of objects (patterns) into a number of predefined classes has been extensively studied in wide variety of applications such as, character recognition, speech recognition and face recognition. These applications often involve hundreds or thousands of classes. To alleviate the burden of classification process, the process is usually divided into two stages: the coarse classification (also called preclassification) process and the fine classification process. To classify an unknown object, firstly, the coarse classification is employed to reduce the large set of candidate objects to a smaller one. Then, the fine classification is applied to identify the class of the object. Generally speaking, we have to extract useful features from the objects (patterns) being classified before the classification process. Thus, we may consider the design of classification systems in terms of two subproblems: (1) feature extraction and (2) classification. The purpose of this paper is to design a general coarse classification scheme, which is low dependent on domain-specific knowledge. To achieve this goal, we need not only reliable and general features in the feature extraction stage, but also general classification method in the coarse classification stage. Therefore, we are motivated to apply statistical approach and a technique which is widely used in image compression, known as Discrete Cosine Transform (DCT) [1], for coarse classification. The DCT helps separate an image into parts of differing importance with respect to the image's visual quality. Due to the energy compacting property of DCT, much of the signal energy has a tendency to lie at low frequencies. Therefore, in our approach only the candidates whose low frequency DCT coefficients close to those of the object being classified will be accepted in the coarse classification process, such that the dimension of features can be largely reduced.

In this paper, we will focus on the topic of improving the efficiency of the classification process. Reliable features of low dimensionality will be extracted for coarse classification. Using these features to perform coarse classification can eliminate a large number of improbable candidates in the early stage and therefore reduce the burden of the subsequent fine classification process. Our coarse classification scheme is based on k-means clustering. On classifying an unknown object, the classes within the clusters that satisfy some thresholds are chosen as candidates. This paper is organized as follows. Section 2 introduces the feature extraction problems. Section 3 explains our coarse classification scheme. Section 4 gives experimental results. Finally, conclusions are drawn in Section 5.

2. Feature Extraction

In general, features are combinations of the measurements that summarize them in a useful way.

In most applications, feature design has not found a general solution which is better than an artisanal solution. Generally speaking, reliable and simple features will be extracted for the coarse classification. Using low dimensional reliable features to perform coarse classification can eliminate a large number of improbable candidates in the early stage and hence reduce the burden of the subsequent fine classification process.

Since the discrimination ability of the features for coarse classification is not as critical as that of the features for fine classification, instead of designing the best features for a certain application, we attempt to find general features that can be applied properly to most vision-oriented applications.

The purpose of this paper is to design a general coarse classification scheme for most vision-oriented applications. To achieve this goal, we need not only reliable and general features in the feature extraction stage, but also general classification method in the coarse classification stage. Therefore, we are motivated to apply DCT to extract statistical features. Then, a statistical-based coarse classification scheme can be developed.

Developed by Ahmed et al. [1], the DCT is a technique for converting a signal into elementary frequency components. Each DCT uses N orthogonal real basis vectors whose components are cosines. The DCT approach has an excellent energy compaction property and requires only real operations in transformation process. The DCT helps separate the image into parts (or spectral sub-bands) of differing importance (with respect to the image's visual quality).

On applying DCT, a frequency spectrum (or the DCT coefficients) $C(u,v)$ of an $N\times N$ image represented by $x(i, j)$ for i, j=0, 1, …, N-1 can be defined as

$$C(u,v) = \frac{2}{N}\alpha(u)\alpha(v)\sum_{i=0}^{N-1}\sum_{j=0}^{N-1}x(i, j) \times \cos(\frac{(2i+1)u\pi}{2N})\cos(\frac{(2j+1)v\pi}{2N}), \quad (1)$$

where

$$\alpha(w) = \begin{cases} \sqrt{1/2} & \text{for } w = 0, \\ 1 & \text{otherwise.} \end{cases}$$

For most images, much of the signal energy lies at low frequencies; these appear in the upper left corner of the DCT. The lower right values represent higher frequencies, and are often small - small enough to be neglected with little visible distortion. Therefore, DCT has superior energy compacting property.

Based on above observations, we were motivated to devise a cluster-based coarse classification scheme using low frequency DCT coefficients as discriminating features. In the next section we shall introduce our coarse classification scheme.

3. Statistical Coarse Classification

There are two distinct philosophies of classification in general, known as "statistical" [2], [3] and "structural" [4]. In the statistical approach the measurements that describe an object are treated only formally as statistical variables, neglecting their "meaning". The structural approach, on the other hand, regards objects as compositions of structural units, usually called *primitives*. Since our main goal is to develop a general coarse classification scheme for vision-oriented applications (such as OCR, face recognition, image retrieval, etc.), the statistical approach is more suitable than the structural approach.

In statistical approach, a pattern is represented by a set of D features viewed as a D-dimensional feature vector. Basically, the statistical classification system is operated in two modes: *training* (*learning*) and *classification* (*testing*). In the training mode, the feature extraction module finds the appropriate features for representing the input patterns and the classifier is trained to partition the feature space. In the classification mode, the trained classifier assigns the input pattern to one of the pattern classes under consideration based on the measured features. To evaluate the performance of a classifier, the *holdout* method [5] can be applied. In the holdout method, the given data are randomly partitioned into two independent sets, a training set and a test set. The training set is used to derive the classifier, whose accuracy is estimated with the test set.

3.1. Techniques to speed up the classification process

In order to speed up the classification process, two techniques can be applied in advance of the time-consuming fine classification process. They are *clustering* and *pruning*.

3.1.1. Clustering

Clustering [4] is the process of grouping the data objects into clusters so that objects within a cluster have high similarity in comparison to one another, but are very dissimilar to objects in other clusters. Its main goal is to identify clusters present in the data. Existing clustering algorithms can be classified into two main categories: hierarchical methods and partitioning methods. Hierarchical methods are either agglomerative (bottom-up) or divisive (top-down). Both suffer from the fact that once a step (merge or split) is done, it can never be undone. In contrast, given the number k of partitions to be found, a partitioning method tries to find the best k partitions of the n objects. Most of the partitioning-based

clustering algorithms adopt one of two popular heuristic methods: (1) the *k*-means algorithm, where each cluster is represented by the mean value of the objects in the cluster, and (2) the *k*-medoids algorithm, where each cluster is represented by one of the objects located near the center of the cluster.

To perform a coarse classification, in the training mode, the feature vectors of learning samples are clustered first based on a certain criterion. In the classification mode, the distances between a test sample and every cluster is calculated, and the clusters that are nearest to the test sample are chosen as candidate clusters. Then the classes within those candidate clusters are selected as the candidates of the test sample.

3.1.2. Pruning

The concept underlying pruning is to eliminate the obvious unqualified candidates via a certain criterion. In the training mode, the template vector of a class is generated from the training samples of the class. In the classification mode, when a test sample is being classified, all the distances between the feature vector of the test sample and every template vector are calculated. Then the templates whose distance to the test sample is beyond a predefined threshold will be pruned. Therefore, for the test sample to be classified, the size of the candidate list is reduced.

3.2. Coarse classification scheme

Before going into the details of our approach, we first briefly review the problem and introduce related symbols used in this paper. The ultimate goal of classification is to classify an unknown pattern x to one of M possible classes ($c_1, c_2,\ldots, c_M$). Each pattern is represented by a set of D features, viewed as a D-dimensional feature vector.

To find the reduced set of candidates for an input pattern, we propose a coarse classification scheme which takes advantage of both clustering and pruning. In our approach the coarse classification is accomplished by four main modules: (1) pre-processing, (2) feature extraction, (3) clustering and (4) pruning.

In the training mode, the training samples are first normalized to a certain size. Then the most significant D features of each training sample are extracted by DCT. For example, the most important 4 features of a pattern are the 4 DCT coefficients with the lowest frequencies, namely $C(0,0)$, $C(0,1)$, $C(1,0)$ and $C(1,1)$. After that the average D features are obtained for each class, assuming there is at least one training sample for each class. Therefore, M feature vectors can be obtained and served as the templates (reference patterns) for each class. After that the *k*-means algorithm [5], [4] is applied for clustering. The *k*-means algorithm first chooses k templates arbitrarily as the initial centers of the k clusters. Then the remaining (M-k) templates are assigned to their nearest clusters one by one, according to a certain distance measure. Each cluster center is updated progressively from the average of the total patterns within the cluster and can be viewed as a virtual reference pattern. The distance between a pattern and a cluster is the distance between the pattern and the representative pattern (i.e., center) of the cluster. In our approach, the sum of squared difference is adopted to evaluate the dissimilarity (distance) between two patterns. Suppose $P_i = [p_{i1}, p_{i2},\ldots, p_{iD}]$ represents the feature vector of pattern P_i. Then the dissimilarity (distance) between patterns P_i and P_j is defined as

$$d(P_i, P_j) = \sum_{d=1}^{D} \left(p_{id} - p_{jd}\right)^2 . \quad (2)$$

After the *k*-means clustering, the M classes of templates are partitioned into k clusters, namely C_1, $C_2,\ldots, C_k$.

In the classification mode, for an input pattern to be recognized, the pattern is first normalized to a certain size. Then the features of the pattern are extracted by DCT. After that the pattern is matched against the center of each cluster to obtain the distance to each cluster. Finally, the clusters which do not satisfy some distance-based thresholds will be excluded from the set of candidate clusters, and the members within the remaining clusters will be served as the candidates for the input pattern. In the next subsection, we will introduce the thresholds used for the coarse classification.

3.2.1. Cluster pruning rules

On classifying a test pattern, the distance between the test pattern x and each cluster C_i is obtained, namely $d(x,C_i)$. To observe the rate of distance per maximum distance, the relative distance $d_m(x, C_i)$ is defined as

$$d_m(x, C_i) = \frac{d(x, C_i)}{Max_{C_k} d(x, C_k)}, \quad (3)$$

where the notation $Max_{C_k} d(x, C_k)$ denotes the maximum distance between x and any cluster. We also define the relative distance $d_a(x, C_i)$ as

$$d_a(x, C_i) = \frac{d(x, C_i)}{Average_{C_k} d(x, C_k)}, \quad (4)$$

where the notation $Average_{C_k} d(x, C_k)$ denotes the average distance over all clusters.

To filter out the clusters which are most dissimilar to the pattern, we devise the following four threshold-

based pruning rules to eliminate unqualified clusters in the coarse classification process.

- **R_1 : Pruning via absolute distance threshold θ_d**

In this rule the distance $d(x,C_i)$ is compared with a pre-determined distance threshold θ_d. If $d(x,C_i)$ is larger than θ_d, then cluster C_i is excluded from the candidate cluster set of x.

- **R_2 : Pruning via cluster rank threshold θ_r**

The aim of this rule is to filter out the farthest clusters by eliminating the clusters whose ranks are larger than θ_r.

- **R_3 : Pruning via relative distance threshold θ_m**

If $d_m(x,C_i)$ is larger than θ_m, then cluster C_i is excluded from the candidate cluster set of x.

- **R_4 : Pruning via relative distance threshold θ_a**

If $d_a(x,C_i)$ is larger than θ_a, then cluster C_i is excluded from the candidate cluster set of x.

Each threshold value is obtained from the statistics of the training samples in the training mode, based on a certain accuracy requirement. Rule R_1 and R_2 are obtained straightforwardly, which cannot measure the similarity between x and each cluster precisely. In contrast, rule R_3 and R_4 take advantage of the discriminating ability of the relative distances. For instance, '$d_a(x,C_i)$=0.5' means the distance between x and cluster C_i is about half those distances between x and other clusters; in other words, cluster C_i is more similar to x than most of the other clusters.

4. Experimental Results

In our application, the objects to be classified are the characters extracted from one of the famous handwritten rare books, Kin-Guan (金剛) bible. Since most of the characters in these rare books were contaminated by various noises, it is a challenge to achieve a high recognition rate. A preliminary experiment has been made to test our approach. There are a total number of 6000 samples (about 500 classes). Each character image was transformed into a 48×48 bitmap. 5000 of the 6000 samples are used for training and the others are used for testing. In our experiment the number of clusters used in k-means clustering is set to 10. The number of DCT coefficients extracted from each sample is set to 9, i.e., $C(i, j)$, where i,j=0, 1, 2.

Table 1 shows the average reduction and accuracy rate of the test samples under the nearest r cluster(s). We can find that high reduction rate always results in low accuracy rate. Table 2 shows the average reduction rate and accuracy rate of the test samples under different rules. From the table we can find that rules R_4, which corresponds to relative distance threshold θ_a, performs better than the other rules.

Table 1. Reduction and accuracy rate under the nearest r cluster(s).

r	1	2	3	4	5	6	7
R.R.	0.90	0.79	0.68	0.57	0.47	0.36	0.26
A.R.	0.635	0.834	0.920	0.958	0.973	0.986	0.994

R.R.: Reduction Rate. A.R.: Accuracy Rate.

Table 2. Reduction and accuracy rate under different pruning rules.

Rule	Exp. 1		Exp. 2		Exp. 3	
	R.R	A.R.	R.R	A.R	R.R	A.R
R_1	0.13	0.99	0.20	0.98	0.26	0.97
R_2	0.26	0.99	0.36	0.986	0.47	0.973
R_3	0.17	0.99	0.21	0.98	0.26	0.97
R_4	0.36	0.99	0.45	0.98	0.50	0.97

R.R.: Reduction Rate. A.R.: Accuracy Rate.

5. Conclusions

This paper presents a coarse classification scheme based on DCT and k-means clustering. The advantages of our approach include:

- Through the energy compacting property of DCT, the extracted features are simple, reliable and appropriate for coarse classification.
- The proposed coarse classification scheme is a general approach to most of the vision-oriented applications.

Future works include the application of another well-known vision-oriented feature extraction method: wavelet transform. Since features of different types complement one another in classification performance, by using features of different types simultaneously, classification accuracy could be further improved.

References

[1] N. Ahmed, T. Natarajan, and K. R. Rao, "Discrete cosine transform," *IEEE Trans. on Comput.*, vol. 23 pp. 90-93, 1974.

[2] T. W. Chiang and T. Tsai, "A Statistical Mask-Matching Approach for Recognizing Handwritten Characters in Chinese Paleography," *in Proc. IEEE Int. Conf. on Systems, Man, and Cybernetics, Hague*, pp. 4717-4721, Oct.2004.

[3] A. K. Jain, P. W. Duin, and J. Mao, "Statistical pattern recognition: a review," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 22, no. 2, pp. 5-37, 2000.

[4] M. Nadler and E. P. Smith, *Pattern Recognition Engineering* Wiley & Sons, Inc., 1993.

[5] J. Han and M. Kamber, *Data Mining: Concepts and Techniques* Academic Press, 2001.