Image Database Classification based on Concept Vector Model

Ruofei Zhang, Zhongfei (Mark) Zhang
Sate University of New York, Binghamton, NY 13902, USA
{rzhang, zhongfei}@cs.binghamton.edu

Abstract

Automatic semantic classification of image databases is very useful for users' searching and browsing, but it is at the same time a very challenging research problem as well. In this paper, we develop a hidden semantic concept discovery methodology to address effective semantics-intensive image database classification. In our approach, each image in the database is segmented into regions associated with homogenous color, texture, and shape features. By exploiting regional statistical information in each image and employing a vector quantization method, a uniform and sparse region-based representation is achieved. With this representation a probabilistic model based on statistical-hidden-class assumptions of the image database is obtained, to which the Expectation-Maximization (EM) technique is applied to analyze semantic concepts hidden in the database. Two methods are proposed to utilize the semantic concepts discovered from the probabilistic model for unsupervised and supervised image database classifications, respectively, based on the automatically learned concept vectors*. It is shown that the* concept vectors *are more reliable and robust than the low level features. The developed methodology has a solid statistical foundation; the theoretic analysis and the experimental evaluations on a database of 10,000 general-purpose images demonstrate its promise of the effectiveness.*

1. Introduction

Automatic image classification is the task of classifying images into semantic categories with or without the supervised training. This categorization of images can be helpful both in the semantic organization of image collections and in obtaining automatic annotations of the images. A common approach to image classification involves addressing the following three issues: (1) image features – how to represent the image; (2) organization of the feature data – how to organize the data; and (3) classifier – how to classify an image. Some work have been reported to address these three issues in the literature. The *configural recognition* scheme proposed by Lipson et al [10] is a knowledge-based scene classification method. A model template, which encodes the common global scene configuration structure using qualitative measurements, is hand-crafted for each category. An image is then classified to the category whose model template best matches the image by deformable template matching. Huang et al [8] proposed a new scheme for automatic hierarchical image classification. Using banded color correlograms, this approach models the features using singular value decomposition (SVD) [4]. Chapelle et al [1] used a trained Support Vector Machine (SVM) to perform image classification. Although shown effective in some specific domains, none of the above techniques ever considers knowledge extracted from the whole image database in the classification. The hidden semantic concept discovery methodology discussed in this paper offers a new approach to classifying image databases into semantic categories with a better effectiveness.

In this paper, we propose new schemes for both supervised and unsupervised automatic image database classification. The schemes are based on the hidden semantic concepts embodied in the database, which are discovered by a probabilistic approach. A new indexing scheme based on a region-image-concept probabilistic model with reasonable assumptions is developed. This model has a solid statistical foundation and is appropriate for the objective of semantics-intensive image database classification. With an iterative Expectation-Maximization (EM) based procedure, the posterior probabilities of each region in an image to hidden semantic concepts are quantitatively obtained, which constitute a semantic concept representation, called *concept vector*, of the image. Based on the obtained *concept vector* representation of each image, two elaborate schemes are developed to classify the image database in unsupervised and supervised manner, respectively. In this way, the effectiveness of the semantic classification in image database is improved because the similarity measure is based on the discovered semantic concepts, which are more reliable and robust than the low-level features used in most existing systems.

2. Concept Model of Image Database

In the proposed approach, the query image and images in the database are first segmented into homogeneous regions. Then representative features are extracted for every region by incorporating color, texture, and shape properties. The image segmentation and corresponding feature extraction method are similar to those employed in [3], which are shown to be effective. Noting that many regions from different images are very similar in terms of the features, a vector quantization (VQ) technique is used to group similar regions together to create a visual dictionary. The visual dictionary for region features is generated by applying Self-Organization Map (SOM) [9] learning (similar idea is used in [15]). SOM is ideal for our problem as it projects high-dimensional feature vectors to a 2-dimensional plane through mapping similar features together while separating different features apart at the same time. Each node in the map represents a region feature set (i.e., a "code word" in the visual dictionary) in which the intra-distance is low. The extent of similarity in each "code word" is controlled by the size of the visual dictionary, which is determined empirically. Based on the visual dictionary, each image can be represented by a uniform vector model. In this representation, an image is a vector with each dimension corresponding to a "code word". Based on this representation of every image, the database is modeled as a $M \times N$ "code word"-image matrix which records the occurrence of every "code word" in each image, where N is the number of images in the database and M is the number of "code words" in the dictionary. In the rest of this paper, we use the terminologies region and "code word" interchangeable; they both denote an entry in the visual dictionary equally.

With a uniform "code word" vector representation for each image in the database, we propose a probabilistic model in a Bayesian framework. We assume that the (region, image) are known i.i.d. samples from an unknown distribution. Furthermore, these samples are associated with an unobserved *semantic concept* variable $z_k \in Z = \{z_1, \ldots, z_K\}$. Each observation of one region ("code word") $r_i \in R = \{r_1, \ldots, r_M\}$ in an image $g_j \in G = \{g_1, \ldots, g_N\}$ belongs to one concept class z_k. To simplify the model, we make two more assumptions. First, observation pairs (r_i, g_j) are generated independently. Second, the pairs of random variable (r_i, g_j) are conditionally independent given the respective hidden concept z_k, i.e., $P(r_i, g_j|z_k) = P(r_i|z_k)P(g_j|z_k)$. These two assumptions are intuitively reasonable.

Following the Maximum Likelihood Estimation (MLE) principle, one determines $P(z_k)$, $P(r_i|z_k)$, and $P(g_j|z_k)$ by maximization of the log-likelihood function

$$\mathcal{L} = \log P(R, G) = \sum_{i=1}^{M} \sum_{j=1}^{N} n(r_i, g_j) \log P(r_i, g_j) \tag{1}$$

where $n(r_i, g_j)$ denotes the number of region r_i occurred in image g_j. From (1) we derive that this is a statistical mixture model [11], which can be resolved by applying the Expectation-Maximization (EM) technique. Applying Bayes' rule with (1), we determine the posterior probability for z_k under (r_i, g_j):

$$P(z_k|r_i, g_j) = \frac{P(z_k)P(g_j|z_k)P(r_i|z_k)}{\sum_{k'=1}^{K} P(z_{k'})P(g_j|z_{k'})P(r_i|z_{k'})} \tag{2}$$

Maximizing the expectation of the complete-data likelihood $\log P(R, G, Z)$ for estimated $P(Z|R, G)$ derived from (2) with Lagrange multipliers to $P(z_l)$, $P(r_u|z_l)$, and $P(g_v|z_l)$, respectively, the parameters are determined as

$$P(z_k) = \frac{\sum_{i=1}^{M} \sum_{j=1}^{N} n(r_i, g_j) P(z_k|r_i, g_j)}{\sum_{i=1}^{M} \sum_{j=1}^{N} u(r_i, g_j)} \tag{3}$$

$$P(r_u|z_l) = \frac{\sum_{j=1}^{N} n(r_u, g_j) P(z_l|r_u, g_j)}{\sum_{i=1}^{M} \sum_{j=1}^{N} u(r_i, g_j) P(z_l|r_i, g_j)} \tag{4}$$

$$P(g_v|z_l) = \frac{\sum_{i=1}^{M} n(r_i, g_v) P(z_l|r_i, g_v)}{\sum_{i=1}^{M} \sum_{j=1}^{N} u(r_i, g_j) P(z_l|r_i, g_j)} \tag{5}$$

Alternating (2) with (3)–(5) defines an iterative procedure that converges to a local maximum of the expectation. For details of the derivation and the technique to determine the number of concepts, K, please refer [15].

3. Concept Vector based Image Classification

Based on the probabilistic model, we can derive the posterior probability of each image in the database to every discovered concept by applying Bayes' rule as

$$P(z_k|g_j) = \frac{P(g_j|z_k)P(z_k)}{P(g_j)} \tag{6}$$

which can be determined with the estimations in (3)–(5). The posterior probability vector $P(Z|g_j) = [P(z_1|g_j), P(z_2|g_j), \ldots, P(z_K|g_j)]^T$ is called the *concept vector* and is used to quantitatively describe the semantic concepts associated with the image g_j. This vector can be considered as a representation of g_j (which originally has a representation in the M-dimensional "code word" space) in the K-dimensional *concept space* determined by the estimated $P(z_k|r_i, g_j)$ in (2).

With the proposed probabilistic model, we are able to concurrently obtain $P(z_k|r_i)$ and $P(z_k|g_j)$ such that both regions and images have an interpretation in the concept space simultaneously, while the image clustering based approaches, e. g. [6], do not have this flexibility. Now every region and/or image can be represented as a weighted sum of the discovered concept axes.

For image database classification, typically two paradigms are applied. One is the unsupervised classification (e.g., [2]) and the other is the supervised classification (e.g., [8]). We develop two simple yet effective classification schemes based on the posterior probabilities of the discovered semantic concepts for the unsupervised and supervised classifications, respectively.

To achieve fast image classification, we develop a hierarchical classification structure for the database and a related algorithm to perform the unsupervised image classification.

Let S denote the set of all the nodes in the classification structure, and X be the set of all images in the database. Each node $s \in S$ is a set of images $X_s \subset X$ with a vector z_s, the centroid of the *concept vector* set $P(Z|x)$ $(x \in X_s)$ in the node. The children of a node $s \in S$ are denoted by $c(s) \subset S$. The child nodes partition the image space of the parent node such that $X_s = \bigcup_{r \in c(s)} X_r$. Now the question is how to construct such an optimal classification structure. We iteratively apply the modified *k*-means algorithm [14] to all the *concept vectors* corresponding to each image in the database to form the hierarchy of the classification structure. All the nodes represent centroid *semantic vectors* of a corresponding set of images. The number of nodes in each level and the depth of the classification structure are determined adaptively based on the iterative threshold parameter in the modified *k*-means algorithm.

Typical search algorithms would traverse the tree top-down, selecting the branch that minimizing the distance between a query q and a cluster centroid z_s . However, this search strategy is not optimal since it does not allow backtracking. To achieve an optimal search, we keep track of all the nodes which have been searched and always select the nodes with the minimum distance to the query region. This search algorithm is guaranteed to select the node whose centroid has the minimum distance in the set of visited nodes to the query region. Hence, it is optimal.

Thus, given a query image, we have the following classification algorithm. The symbols used in the algorithm are introduced below: s^* is the node whose centroid has the minimum distance to the query *concept vector* q; ts is the threshold of the size of a node that q is classified to; Ω is the node set we have searched; $|c(s^*)|$ is the size of the child set of s^*; z_s is the node centroid; *NodesSearched* records the number of nodes we have searched so far; $DIST(\bullet)$ is the distance metric used in the algorithm. The resulting Ψ is the image set to which the query image is classified.

```
input  : $q$, the query image
input  : $ts$, the size threshold
output : $\Psi$, the node that $q$ is classified to
begin
    $s^* = root$;
    $\Omega = \{s^*\}$;
    $NodesSearched = 0$;
    while $\|s^*\| > ts$ do
        $\Omega \leftarrow (\Omega - \{s^*\}) \cup c(s^*)$;
        $NodesSearched = NodesSearched + |c(s^*)|$;
        $s^* \leftarrow \arg\min_{s \in \Omega}(DIST(q, z_s))$;
    end
end
```

Algorithm 1: The unsupervised classification algorithm.

For the supervised classification problem, with the *concept vector* of each image in the training set, we build a classification tree by applying C4.5 algorithm [5] on the *concept vector* set. We assume that each image in the training set belongs to only one semantic category. The splitting attribute selection for each branch is based on the information gain ratio [13]. Associated with each leaf node of the classification tree is a ratio m/n, where n is the number of images classified to this node and m is the number of incorrectly classified images. This ratio is a measure of the classification inaccuracy of the classification tree for each category in the training image set.

The training set used to test the *concept vector* based supervised classification method consists of 10 fairly representative categories of the COREL images (40 images in each category); the 10 image categories are: *African people (a1), beach (a2), medieval buildings (a3), buses (a4), dinosaurs (a5), elephants (a6), flowers (a7), horses (a8), mountains and glaciers (a9)*, and *European dishes (a10)*. These images contain a wide range of content (scenery, animal, objects, etc.). The classification tree built is shown in Fig. 1.

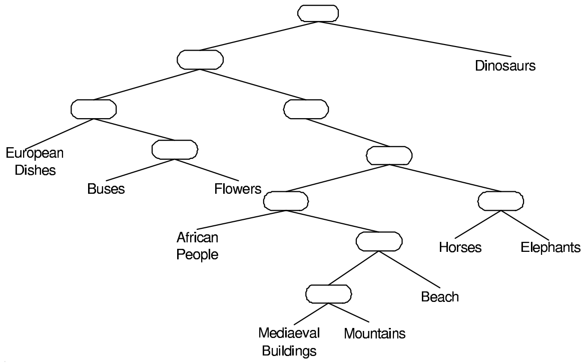

Figure 1: The classification tree obtained from a training set.

4. Experiment Results

We have implemented the approach in a prototype system on a platform of Pentium IV 2.0 GHZ CPU and 512M memory. The following reported evaluations are performed on a general-purpose color image database containing 10,000 images from the COREL collection with 96 semantic categories. These categories include *landscape, fashion, historical building, city life,* etc. Each semantic category consists of 85–120 images. In the case of evaluating supervised classification, the 10,000 images are partitioned to a training set and a testing set. The training set is composed of half number of images from each category and all the remaining images constitute the testing set.

In the experiment, the parameters of the image segmentation algorithm [14] is adjusted considering the balance of the depiction detail and the computation intensity such that there are in average 8.3207 regions in each image. To determine the size of the visual dictionary, different numbers of "code words" have been selected and the *average classification accuracy* of the classification tree built for the training set has been evaluated. Two statistics of the classification performance are recorded for the testing and training sets. They are *Average classification error rate*: The average rate that a query image is misclassified, and *Average classification accuracy*: The average value of the classification accuracy for training images in all categories (the average value of $(1 - m/n)$ described in Sec. 3).

The average classification accuracy and the average classification error rate vs. the number of "code words" in the visual dictionary is shown in Fig. 2. It is indicated that the general trend is that the larger the visual dictionary size, the higher the classification accuracy and the lower the classification error rate. However, a larger visual dictionary size means a larger number of image feature vectors, which implies a higher computation complexity in the hidden semantic concept discovery. Also, a larger visual dictionary leads to a larger storage space. Therefore, we use 800 as the number of the "code words", which corresponds to the first turning point for both the classification accuracy and the classification error rate curves in Fig. 2. Since there are in total 83,307 regions in the database, in average each "code word"

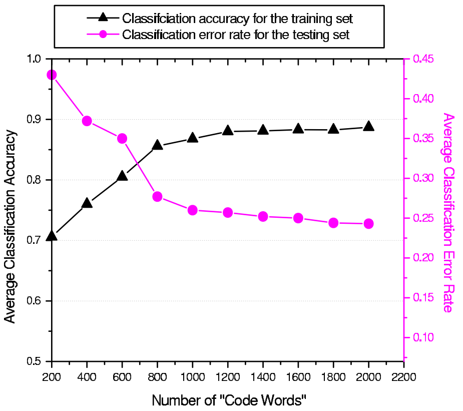

Figure 2: Average classification accuracy for different sizes of the visual dictionary.

Table 1: Average relevancy ratios for the 500 queries in color variations by using *concept vectors* and banded color correlograms.

Average Relevancy Ratio	Color percentile variation(%)				
	0	5	10	15	20
color correlograms	0.771	0.740	0.630	0.594	0.483
concept vectors	0.878	0.869	0.840	0.832	0.807

represents 104.13 regions.

Following the principle of Minimum Description Length (MDL) (details can be found in [15]), the number of the concepts is determined to be 132. Performing the EM model fitting, we have obtained the conditional probability of each “code word” to every concept, i. e., $P(r_i|z_k)$. In terms of the computational complexity, despite of the iterative nature of EM, the computing time for the model fitting at $K = 132$ is acceptable (less than 1 second). The average number of iterations upon convergence for one image is less than 5.

The hierarchical unsupervised classification scheme described in Section 3 for the 10,000-image COREL database is constructed. To evaluate the performance of the scheme and the related classification algorithm , 500 images are randomly selected from all the categories as the query set. The ratios of the relevant images in the node returned by the classification algorithm (relevancy ratios) are subjectively examined by users. The reliability and robustness of the derived *concept vectors* for improving the unsupervised classification accuracy are evaluated. The performances of *concept vector* and banded color correlgorams [7] for different degrees of color variations are compared by applying the same scheme and classification algorithm. Color variations can be simulated by changing colors to their adjacent values for each image. We apply color changes to an query image, then the modified image is used as the query image, and the ratio of relevant images in the returned node is recorded. The average relevancy ratios of the 500 queries in different color variations are recorded in Table 1. The experiment reports that *concept vectors* are more reliable and robust than the color correlgorams; the performance of *concept vectors* is much higher (more than 10%) than that of the banded color correlgorams due to the improved reliability. At the same time, the performance of *concept vectors* decreases gracefully when the color variation level increases while the color correlograms are much more sensitive to the color variations.

To provide quantitative evaluations on the performance of the supervised image classification, we run the prototype on a controlled subset of the COREL collection. This controlled database consists of 10 image categories the same as the training set *a1* to *a10* described in Section 3, each containing 100 pictures. Within this controlled database, we can assess classification performance reliably with the ground-truthed categorization information because the categories are semantically non-ambiguous and share no semantic overlaps.

The classification performance of the constructed classification tree is compared with the classification method developed by Huang et al [8]. In Huang et al’s method, the banded color correlograms [7] are used as the features extracted. For both methods, 40 randomly selected images for each category are used to train the classifiers; the classification methods are then tested using the rest 600 images outside the training set. The classification results of our proposed method and the normalized cuts based classification method [8] are shown in Table 2. In both tables each row lists the percentage of images in one category classified to each of the 10 categories. Numbers in the diagonal show the classification accuracy for every category. The classification behavior of our proposed method is clearly better than that of the normalized cuts based method since (i) the overall number of misclassifications between categories is smaller and (ii) the overall number of correct classifications is larger. The average classification error rate of our method is lower than that of Huang et al’s method by 12.8%.

5. Conclusions

This paper is about automatic general-purpose image database classification. The main contributions of this work are the identification of the problems existing in most existing methods —- unreliable feature evidence on semantic contents, and the development of classification methods based on more semantics-sensitive features to solve for the problems. Performing image segmentation with multiple features and developing a SOM based quantization method to generate a visual dictionary, a uniform and sparse region-based representation scheme is obtained. On the basis of this representation a probabilistic model of the image database is defined. Based on this model, a EM-based procedure is applied to discover the hidden semantic concepts in the database. Two methods are proposed to utilize the semantic concepts discovered from the probabilistic model for unsupervised and supervised image database classifications, respectively, based on the automatically learned *concept vectors*. Supported by the solid statistical foundation, this approach enables a representation by higher order se-

Table 2: Results of the discovered semantic concepts based (upper) and the normalized cuts based image classification experiments for the controlled database.

%	a1	a2	a3	a4	a5	a6	a7	a8	a9	a10
a1	**40**	0	1	0	4	8	5	0	2	0
a2	0	**28**	2	0	0	0	1	1	28	0
a3	3	1	**47**	0	3	2	0	0	2	2
a4	0	9	2	**37**	0	4	0	1	6	1
a5	0	0	0	0	**60**	0	0	0	0	0
a6	2	0	0	0	2	**41**	0	2	13	0
a7	0	0	1	0	0	0	**54**	0	1	4
a8	0	0	0	1	0	2	11	**39**	3	4
a9	0	4	4	0	1	1	0	0	**50**	0
a10	4	2	0	1	3	0	5	0	4	**41**

%	a1	a2	a3	a4	a5	a6	a7	a8	a9	a10
a1	**29**	6	4	0	2	6	3	4	5	1
a2	1	**30**	1	0	0	9	0	7	7	5
a3	4	4	**27**	2	3	10	0	2	8	0
a4	1	7	5	**32**	0	3	0	1	10	1
a5	0	0	1	0	**52**	0	4	3	0	0
a6	2	0	3	0	1	**37**	0	4	10	3
a7	1	1	1	5	0	0	**45**	0	1	6
a8	0	1	0	2	0	3	6	**38**	5	5
a9	2	5	5	0	2	2	0	0	**41**	3
a10	5	1	2	3	1	3	5	0	11	**29**

mantic indicants which are more reliable and robust, hence improves the image classification accuracy. The experimental evaluations on a database of 10,000 general-purpose images demonstrate the effectiveness and the promise of the approach in both supervised and unsupervised image classifications.

References

[1] O. Chapelle, P. Haffner, and V. N. Vapnik. Support vector machines for histogram-based image classification. *IEEE Trans. on Neural Networks*, 10(5):1055–1064, September 1999.

[2] A. Chardin and P. Perez. Unsupervised image classification with a hierarchical em algorithm. In *IEEE International Conference on Computer Vision*, Corfu, Greece, September 1999.

[3] Y. Chen and J. Z. Wang. A region-based fuzzy feature matching approach to content-based image retrieval. *IEEE Trans. on PAMI*, 24(9):1252–1267, 2002.

[4] S. Deerwester, S. Dumais, G. Furnas, T. Landauer, and R. Harshman. Indexing by latent semantic analysis. *Journal of American Sociation of Information Science*, 41:391–407, 1990.

[5] M. H. Dunham. *Data Mining, Introductory and Advanced Topics*. Prentice Hall, Upper Saddle River, NJ, 2002.

[6] F. J. et al. An efficient region-based image retrieval framework. In *ACM Multimedia Proceedings*, Juan-les-Pins, France, December 2002.

[7] J. Huang and S. R. K. et al. Image indexing using color correlograms. In *IEEE Int'l Conf. Computer Vision and Pattern Recognition Proceedings*, Puerto Rico, 1997.

[8] J. Huang, R. Kumar, and R. Zabih. An automatic hierarchical image classification scheme. In *The Sixth ACM Int'l Conf. Multimedia Proceedings*, 1998.

[9] T. Kohonen, S. Kaski, K. Lagus, J. Salojärvi, J. Honkela, V. Paatero, and A. Saarela. Self organization of a massive document collection. *IEEE Trans. on Neural Networks*, 11(3):1025–1048, May 2000.

[10] P. Lipson, E. Grimson, and P. Sinha. Configuration based scene classification and image indexing. In *The 16th IEEE Conf. on Compuer Vision and Pattern Recognition Proceedings*, pages 1007–1013, 1997.

[11] G. Mclachlan and K. E. Basford. *Mixture Models*. Marcel Dekker, Inc., Basel, NY, 1988.

[12] J. Rissanen. *Stochastic Complexity in Statistical Inquiry*. World Scientific, 1989.

[13] C. Shannon. Prediction and entropy of printed english. *Bell Sys. Tech. Journal*, 30:50–64, 1951.

[14] J. Z. Wang, J. Li, and gio Wiederhold. Simplicity: Semantics-sensitive integrated matching for picture libraries. *IEEE Trans. on PAMI*, 23(9), September 2001.

[15] R. Zhang and Z. Zhang. Hidden semantic concept discovery in region based image retrieval. In *IEEE International Conference on Computer Vision and Pattern Recogntion (CVPR) 2004*, Washington, DC, June 2004.