Multi-category SVMs-based Image Categorization

Yutao Han and Xiaojun Qi

Computer Science Department, Utah State University, Logan, UT 84322-4205
yhan@cc.usu.edu and xqi@cc.usu.edu

Abstract

Automatic image categorization, which maps low-level visual features to high-level semantics, is the crucial basis for effective understanding, annotation, retrieval, and management of digital visual information. In this paper, we present a multi-category SVMs-based (Support Vector Machines) image categorization system by exclusively using global low-level features. Images are represented by the MPEG-7 SCD (Scalable Color Descriptor) and the modified MPEG-7 EHD (Edge Histogram Descriptor), which are verified by the K-Means algorithm to be effective. The K-Means algorithm is also applied to select training images. Multi-category SVMs using the Gaussian RBF (Radial Basis Function) kernels are further evaluated on the categorization of the image database with general-purpose images from COREL. The empirical results demonstrate that the proposed categorization method outperforms the peer approaches in the literature in terms of both efficiency and accuracy. In addition, each test image is labeled by a set of confidence values for each possible category.

Keywords: Image categorization, support vector machines, MPEG-7 color and texture descriptors.

1. Introduction

Image categorization is to label images into one or several predefined categories. It is mainly used for visual information management and can be applied in a variety of domains such as entertainment, education, biomedicine, web image classification and search, etc. In particular, it can aid in retrieval since categorized keywords greatly narrow the semantic gaps.

Automatic image categorization is a challenging task due to various imaging conditions, complex and hard-to-describe objects, highly textured background, and occlusions. In general, most approaches use learning-based techniques to train manually categorized training images and test the uncategorized images based on the training results. The images are represented by either global color histograms [1, 2], or block-based local and spatial properties [3, 4], or region-based local features [5-9]. Several generic image categorization systems are briefly reviewed.

Huang et al. [1] construct a color-correlogram-based classification tree for image categorization. Chapelle et al. [2] apply SVMs on the global 16×16×16-bin HSV color histograms to categorize images. Li and Wang [3] propose an ALIP system which uses the 2D multi-resolution hidden Markov model on color and texture features of image blocks of size 4x4 for image classification. Murphy et al. [4] build four graphical models to relate features of image blocks to objects and perform joint scene and object recognition. Modestino and Zhang [5] use a Markov random field model to capture spatial relationships between regions and apply a maximum *a posterior* rule to interpret images. Smith and Li [6] classify images by applying a composite region template descriptor matrix on the spatial orderings of regions, whose attributes are represented by symbols in a finite pattern library. Barnard and Forsyth [7] apply a hierarchical statistic model to generate keywords for classification based on a sequence of semantically meaningful regions. Jeon *et al.* [8] use the cross media relevance model to predict the probability of generating a word given the regions in an image. Andrews et al. [9] propose an MI-SVM (Multiple Instance) approach to automatic image classification.

In spite of their successes, all these categorization systems have their shortcomings. First, these systems cannot accurately represent each object in an image by features. For example, global color histograms cannot precisely represent the objects which correspond to the semantic contents of an image. Block-based features often break an object into several blocks or put different objects into a single block. Region-based features have the same problems as the block-based features due to inaccurate segmentation. Second, they assign keyword(s) to each image without any weight.

In this paper, we propose a multi-category SVMs-based categorization system to classify the generic images by the likelihood of each predefined category. The remainder of the paper is organized as follows. Section 2 describes our proposed approach. Section 3 illustrates the experimental results. Section 4 draws conclusions.

2. Proposed Approach

2.1. Low-level feature extraction

Since finding accurate representation of an image is an unsolved open research area, we exclusively use global features. In general, global features are easy to extract and have been widely used in retrieval and categorization. In our system, we combine MPEG-7 color and texture features to represent an image.

The SCD is one of the four MPEG-7 normative color descriptors [10]. It uses the HSV color histograms to represent an image since the HSV color space provides an intuitive representation of color and approximates human's perception. We directly adopt the SCD to extract color related information.

The EHD is one of the three normative texture descriptors used in MPEG-7 [10]. It captures the spatial distribution of edges in an image and has been proven to be useful for image retrieval. Five types of edges, namely, vertical, horizontal, 45° diagonal, 135° diagonal, and non-directional, have been used to represent the edge orientation in 16 non-overlapping sub-images. The normative EHD is therefore a total of 5×16 histogram bins. Based on the EHD, we construct gEHD (global EHD) and sEHD (semi-global EHD) to address any segmentation and rotation, scaling, and translation related issues. The gEHD represents the edge distribution for the entire image and has five bins. For the sEHD, we group connected sub-images into 13 different clusters (Fig. 1) and construct the EHD for each cluster. The sEHD and gEHD are used as our final texture feature representation. As a result, the problems associated with inaccurate image segmentation or sub-blocking are avoided.

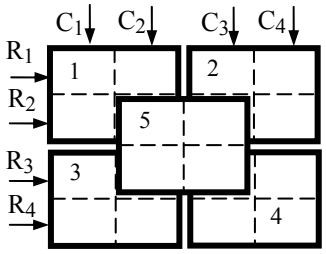

Fig 1: Sub-image clusters for sEHD

2.2. Support vector machines

The SVMs-based approach is considered as a good learning-based candidate due to its strong theoretical foundations, its high generalization performance, and its excellent empirical successes in many applications. The SVMs are designed for the binary classification. In the simplest form, they are hyperplanes that separate training data by a maximal margin. That is, given *m* training data $\{x_i, y_i\}$'s, where $x_i \in R^n, y_i \in \{-1,1\}$, SVMs need to solve:

$$\min_{\omega,b,\xi} \quad \frac{1}{2}\omega^T\omega + C\sum_{i=1}^{l}\xi_i \qquad (1)$$

$$\text{Subject to } \; y_i(\omega^T\phi(x_i)+b) > 1-\xi_i, \xi_i > 0$$

where *C* is the penalty parameter of the error term and $K(x_i, y_i) = \phi(x_i)^T\phi(x_j)$ is the kernel function. The SVMs with the RBF kernel are used in our system since they give excellent results compared to linear and polynomial kernels trained with the SVMs and RBFs trained with the non-SVMs [13]. The RBF kernels are:

$$K(x_i, y_i) = \exp\left(\gamma\left\|x_i - x_j\right\|^2\right), \gamma > 0 \qquad (2)$$

As a result, two SVMs related parameters *C* and γ need to be predetermined for the image categorization. We combine the 3-fold cross-validation and grid-search algorithms to find the best *C* and γ for our image database by testing on exponentially growing sequences of $C = 2^{-5}, 2^{-3}, \cdots, 2^{15}$ and $\gamma = 2^{-15}, 2^{-13}, \cdots, 2^{3}$. Experimental results show that C=8 and γ=0.125 achieve the best categorization accuracy.

Furthermore, an appropriate multi-class method is needed when dealing with several classes as in image classification. We use "one against one" approach [11] (i.e., apply pairwise comparisons between classes) in our system due to its better categorization accuracy. In addition, we also map the SVM outputs into probabilities by training the parameters of an additional sigmoid function [12]. Consequently, our system returns the likelihood of each category that an image may belong to.

2.3. Training sample selection

To our knowledge, the training images are randomly selected in other learning-based categorization systems. However, we found that the categorization accuracy can be improved by selectively choosing the training images.

First, we apply the K-Means algorithm to group the images in each predefined category into 3 clusters as shown in Fig. 2. The 3 clusters roughly accommodate all the possible variations in each category in our database, especially in the image category with highly textured objects or backgrounds.

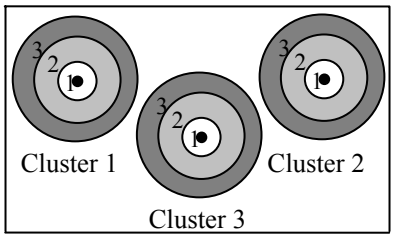

Fig. 2: Training image selection for each category

Second, we choose the training images based on the clustering results of each predefined image category. Several selection strategies have

been compared to determine the one with the best categorization accuracy. These selection strategies are:

1. S1: Randomly choose images with small distances to the cluster centers (i.e., region 1 in each cluster in Fig. 2);
2. S2: Randomly choose images with large distances to the cluster centers (i.e., region 3 in each cluster in Fig. 2);
3. S3: Randomly choose images with median distances to the cluster centers (i.e., region 2 in each cluster in Fig. 2);
4. S4: Randomly choose images with small and large distances to the cluster centers (i.e., regions 1 and 3 in each cluster in Fig. 2).

3. Experimental Results

The proposed image categorization approach has been validated by testing with a general-purpose image database including about 60,000 images. To provide numerical results on the performance, we evaluate the system based on a controlled subset of the COREL database. This subset contains 10 semantically distinct categories with 100 images in each category.

3.1. Feature selection

To verify the effectiveness of the extracted low-level features, the K-means algorithm is used to cluster the feature vectors of training images into 10 image categories (clusters). The effectiveness of the low-level features is measured by the number of images that are correctly clustered. Obviously, the most effective low-level features should group all the training images in the same category into one cluster.

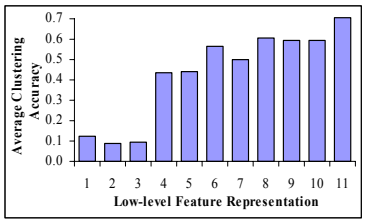

Fig. 3: Clustering results of different features

Several SCD and EHD based features have been tested on several sets of training images which are randomly chosen from each category. The clustering-based effectiveness of the tested low-level features is shown as the average clustering accuracy in Fig. 3, where the low-level features are: (1) 16-bin SCD, (2) 32-bin SCD, (3) 64-bin SCD, (4) sEHD, (5) gsEHD (gEHD+sEHD), (6) 16-bin SCD+sEHD, (7) 16-bin SCD+gsEHD, (8) 32-bin SCD+ sEHD, (9) 32-bin SCD+ gsEHD, (10) 64-bin SCD +sEHD, and (11) 64-bin SCD+gsEHD. It is clearly observed that the last feature is the most effective one since it yields the highest clustering accuracy. The length of this feature is 134 (64 for SCD; 5×13 for sEHD; and 5 for gEHD), which is very compact compared to features used in other systems (i.e., 4096 in [2] and 6144 in [3]).

3.2. Training image selection and categorization

To measure the effectiveness of each proposed training image selection scheme, we choose training images from each category by using the random selection scheme (R1) and four selection strategies introduced in section 2.3. In our experiment, 40 images from each category are chosen as the training images and the remaining 600 images are used as the testing images. The overall categorization accuracy for each selection scheme is shown in Table 1. It clearly shows that the strategy S4 outperforms other strategies, and therefore is implemented in our system. The categorization accuracy improvement is: S4 vs. R1: 3.8%, S4 vs. S1: 18.8%, S4 vs. S2: 5.1%, S4 vs. S3: 10.8%.

Table 1: Overall accuracy of different schemes

Methods	R1	S1	S2	S3	S4
Ave. Accuracy	0.79	0.69	0.78	0.74	0.82

The proposed system is compared with ALIP [3] and our implemented HistSVM [2]. The overall categorization accuracy of ALIP, HistSVM, and our systems on the same controlled subset is 63.6%, 75.3%, and 82.2%, respectively. Our system performs better than the ALIP system with an 18.6% difference in the overall accuracy without using any complicated segmentation and statistical approaches as in [3]. Our system also improves the accuracy by 9.2% over the HistSVM system by using efficient features whose length is roughly 1/30 of theirs.

Fig. 4 plots the average categorization accuracy for each predefined image category by applying the proposed multi-category SVMs-based method (Prop. 1), a variant of the proposed method where random training image selection is involved (Prop. 2), the HistSVM system [2], and the ALIP system [3]. It clearly illustrates that the proposed system achieves the best average accuracy in most categories. In particular, the variant of our proposed approach also yields better average accuracy in most categories than the ALIP and HistSVM systems, where the same random selection scheme is used for choosing the training images. The confusion matrices of ALIP and our systems, which list the detailed classification errors, are also shown in Table 2 in the form of X/Y, where X's are the ALIP's results and Y's are ours.

A few categorization examples are shown in Fig. 5, where the automatically labeled keywords with the top two confidence values are listed below each image.

Table 2: Confusion matrix of the proposed system and the ALIP system, where each row lists the average percentage of the images in one category classified into each of the 10 categories. Numbers on the diagonal show the categorization accuracy for each category.

	Africa	beach	Build.	buses	dinosaur	elephant	flower	horse	Mount.	food
Africa	**52/82**	2/0	4/3	0/3	8/0	16/7	10/0	0/0	6/0	2/5
beach	0/0	**32/72**	6/3	0/2	0/0	0/3	2/0	2/0	58/20	0/0
building	8/20	4/2	**64/55**	0/3	8/0	6/2	0/0	0/2	6/0	4/5
buses	0/0	18/2	6/2	**46/95**	2/0	8/2	0/0	0/0	16/0	4/0
dinosaur	0/0	0/0	0/0	0/0	**100/98**	0/0	0/0	0/0	0/2	0/0
elephant	8/13	0/0	2/2	0/0	8/0	**40/63**	0/0	8/5	34/5	0/12
flower	0/7	0/0	2/0	0/0	0/0	0/0	**90/93**	0/0	2/0	6/0
horse	0/0	2/0	0/0	0/0	0/0	4/2	24/0	**60/97**	4/0	6/2
mountain	0/2	6/13	6/0	0/0	2/0	2/7	0/0	0/0	**84/78**	0/0
food	6/3	4/3	0/0	2/0	6/0	0/5	8/0	0/0	6/0	**68/88**

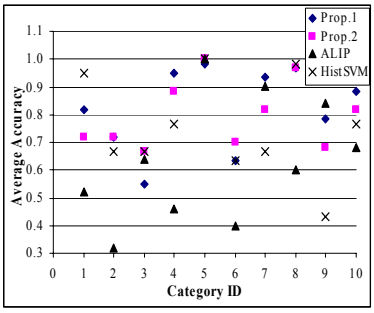

Fig. 4: Comparisons of the average categorization accuracy by using four different methods

Horse: 96% Food: 1% | Building: 92% Beach: 3% | Beach: 47% Mountain: 40% | Buses: 31% Building: 25%

Fig. 5: Categorization examples

4. Conclusions

In this paper, we present an efficient and effective image categorization system. The main contributions are: 1) Represent an image by efficient MPEG-7 based color and texture features. 2) Randomly choose the training images from each category by the K-Means algorithm so more varieties of the images are accommodated. 3) Apply multi-category SVMs to classify images by a set of confidence values for each possible category.

The proposed system has been validated by testing with 60,000 general-purpose images. The controlled experiment has been performed on 10 semantically distinct categories. The experimental results indicate that our system achieves the best overall categorization accuracy compared to two peer systems.

The proposed system can be easily integrated into the retrieval system, where categorized keywords and the query image(s) can be combined as the query. Furthermore, user's relevance feedback can be added to dynamically update the categorized images so the categorization accuracy can be improved.

In the future, we will investigate other low-level features and machine learning algorithms for image categorization. In addition, we will incorporate this promising image categorization system into our image retrieval system to improve the retrieval accuracy.

5. References

[1] J. Huang, S. R. Kumar, and R. Zabih, "An Automatic Hierarchical Image Classification Scheme," *Proc. of 6th ACM Int'l Conf. on Multimedia*, pp. 219-228, 1998.

[2] O. Chapelle, P. Haffner, and V. N. Vapnik, "Support Vector Machines for Histogram-Based Image Classification," *IEEE Trans. on Neural Networks*, vol. 10, no. 5, pp. 1055–1064, 1999.

[3] J. Li and J. Z. Wang, "Automatic Linguistic Indexing of Pictures by a Statistical Modeling Approach," *IEEE Trans. on PAMI*, vol. 25, no. 10, pp. 1075-1088, 2003.

[4] K. Murphy, A. Torralba, and W. Freeman, "Using the Forest to See the Trees: A Graphical Model Relating Features, Objects, and Scenes," *Advances in Neural Information Processing Systems*, vol. 16, 2004.

[5] J. W. Modestino and J. Zhang, "A Markov Random Field Model-Based Approach to Image Interpretation," *IEEE Trans. on PAMI*, vol. 14, no. 6, pp. 606-615, 1992.

[6] J. R. Smith and C. S. Li, "Image Classification and Querying Using Composite Region Templates," *Int'l J. Computer Vision and Image Understanding*, vol. 75, no. 1-2, pp. 165-174, 1999.

[7] K. Barnard and D. Forsyth, "Learning the Semantics of Words and Pictures," *Proc. of Int. Conf. on Computer Vision*, vol. 2, pp. 408-415, 2001.

[8] J. Jeon, V. Lavrenko and R. Manmatha, "Automatic Image Annotation and Retrieval Using Cross-Media Relevance Models," *Proc. of the 26th Intl. ACM SIGIR Conf.*, pp. 119–126, 2003.

[9] S. Andrews, I. Tsochantaridis, and T. Hofmann, "Support Vector Machines for Multiple-Instance Learning," *Advances in Neural Information Processing Systems 15*, pp. 561–568, 2003.

[10] B. S. Manjunath, P. Salembier, and T. Sikora, *Introduction to MPEG-7 Multimedia Content Description Interface*, John Wiley & Sons, 2002.

[11] T. Wu, C. Lin, and R. Weng, "Probability Estimates for Multi-class Classification by Pariwise Coupling", *Journal of Machine Learning Research*, vol. 5, pp. 975-1005, 2004.

[12] J. C. Platt, "Probabilistic Output for Support Vector Machines and Comparisons to Regularized Likelihood Methods," A Bartlett, P Schölkopf, B Schuurmans, E eds. *Advances in Large Margin Classifiers*, MIT Press Cambridge, MA, pp. 61-74, 2000.

[13] B. Scholkopf, K. Sung, C. Burges, F. Girosi, P. Niyogi, T. Poggio, and V. Vapnik, "Comparing Support Vector Machines with Gaussian Kernels to Radial Basis Function Classifiers," *A.I. Memo 1599*, MIT 1996.