Machine-Learning-Based Image Categorization

Yutao Han and Xiaojun Qi

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

Abstract. In this paper, a novel and efficient automatic image categorization system is proposed. This system integrates the MIL-based and global-feature-based SVMs for categorization. The IPs (Instance Prototypes) are derived from the segmented regions by applying MIL on the training images from different categories. The IPs-based image features are further used as inputs to a set of SVMs to find the optimum hyperplanes for categorizing training images. Similarly, global image features, including color histogram and edge histogram, are fed into another set of SVMs. For each test image, two sets of image features are constructed and sent to the two respective sets of SVMs. The decision values from two sets of SVMs are finally incorporated to obtain the final categorization results. The empirical results demonstrate that the proposed system outperforms the peer systems in terms of both efficiency and accuracy.

1 Introduction

Automatic image categorization has become more and more important with the development of Internet and the growth in the size of image databases. Finding relevant images from Internet and a large size image database is not a trivial task if images are not annotated. Manual categorization is a possible solution, but it is time-consuming and subjective. As a result, many researchers have focused on automatic image categorization. A few existing systems are briefly reviewed here.

Huang *et al.* [1] categorize images by using a classification tree, which captures the spatial correlation of colors in an image. Chapelle *et al.* [2] apply SVMs on the global 16×16×16-bin HSV color histograms to categorize images. Smith and Li [3] classify images by applying a composite region template descriptor matrix on the spatial orderings of regions. Barnard and Forsyth [4] apply a hierarchical statistic model to generate keywords for classification based on semantically meaningful regions. Jeon *et al.* [5] use the cross media relevance model to predict the probability of generating a word given the regions in an image. Li and Wang [6] propose an ALIP system which uses the 2D multi-resolution hidden Markov model on features of image blocks for classification. Murphy *et al.* [7] build 4 graphical models to relate features of image blocks to objects and perform joint scene and object recognition.

Recently, MIL (Multiple Instance Learning) has been applied for automatic image categorization. Maron and Ratan [8] use the DD (Diverse Density) learning algorithm for natural scene classification. Zhang and Goldman [9] use EM-DD algorithm, which combines EM (Expectation Maximization) with DD, to achieve a fast and scalable categorization. Andrews *et al.* [10] propose an MI-SVM approach for categorization,

M. Kamel and A. Campilho (Eds.): ICIAR 2005, LNCS 3656, pp. 585 – 592, 2005.

where region-based image features are iteratively fed into SVMs until there are no updates in positive images. Chen and Wang [11] use the DD-SVM method, which combines EM-DD with SVMs for image categorization. Experimental results [10, 11] show that DD-SVM achieves the best categorization accuracy.

In spite of their successes, all these categorization systems have their shortcomings. Global-feature-based systems [1, 2] cannot precisely represent the semantics of an image, which corresponds to objects. Region-based systems [3-5] often break an object into several regions or put different objects into a single region due to inaccurate image segmentation. The block-based [6, 7] and MIL-based systems [8-11] have the similar problems as the region-based systems.

In this paper, we propose a novel machine-learning based approach, which combines MIL-based and global-feature-based SVMs, for image categorization. The MIL-based SVMs apply MIL on the segmented images to find the IPs (Instance Prototypes). The IPs-based image bag features are further used as inputs to a set of SVMs to find the optimum hyperplanes. To address the inaccurate segmentation issues, we create the global-feature-based SVMs, where MPEG-7 SCD (Scalable Color Descriptor) and the modified MPEG-7 EHD (Edge Histogram Descriptor) are used as the global features. For each test image, two sets of image features are constructed and sent to the two respective sets of SVMs. The decision values from two sets of SVMs are finally incorporated to obtain the final categorization results.

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 MIL-Based SVMS

Image Segmentation. To segment an image into coherent regions, the image is first divided into non-overlapping blocks of size 2×2 and a color feature vector (i.e., the mean color of the block) is extracted for each block. The Luv color space is used because the perceptual color difference of the human visual system is proportional to the numerical difference in this space.

After obtaining the color features for all blocks, an unsupervised K-Means algorithm is used to cluster these color features. This segmentation process adaptively increases the number of regions C (initially set as 2) until two termination criteria are satisfied. That is: (1) the total distance D_i from each block to the corresponding cluster center in the i^{th} iteration is less than T_1; or (2) the absolute difference between the total distances of the current and previous iterations (i.e., $|D_i - D_{i-1}|$) is less than T_2. These two thresholds are empirically chosen so reasonable segmentation can be achieved on all images in our test database.

Based on the segmentation results, the representative color feature $\vec{f}_j^{\,c}$ for each region j is calculated by the mean of color features of all the blocks in region j. The representative texture feature $\vec{f}_j^{\,t}$ for each region j is computed by the average energy in each high frequency band after 2-level wavelet decompositions. The wavelet transformation is applied to a "texture template" image obtained by keeping all the pixels

in region *j* intact and setting all the pixels outside region *j* as white. The length of the feature vector for each region is 9 with 3 color features and 6 texture features.

Multiple-Instance Learning (MIL). MIL was originally studied by Dietterich *et al.* [12] in drug activity prediction and has recently received much attention in machine learning. In MIL, each image is a bag and its segmented regions are instances. Its objective is to find the commonalities in all positive images given a set of labeled images. The EM-DD [9] method solves this problem by finding the maximum *DD* value at point *t* in the bag feature space:

$$DD(t) = \arg\max_{t} \prod_{i}^{n} \Pr(B_i, l_i \mid t) \tag{1}$$

where B_i is the i^{th} bag, l_i is the label of the i^{th} bag, and n is the total number of labeled bags. This maximum *DD* value indicates a higher probability that point *t* fits better with the instances from positive bags than those from negative bags. The negative log transformation can be further used to simplify (1):

$$NLDD(t) = \arg\min_{t} \sum_{i=1}^{n} \left(-\log(\Pr(l_i \mid t, B_i))\right) = \arg\min_{t} \sum_{i=1}^{n} \left(-\log(1 - \left|l_i - Label(B_i \mid t)\right|)\right) \tag{2}$$

where $Label(B_i \mid t) = \max_{j} \left\{ \exp\left[-\sum_{d=1}^{n} (s_d (B_{ijd} - t_d))^2 \right] \right\}$, B_{ij} is the j^{th} instance of bag *i*, and S_d refers to the feature weight on dimension *d*. That is, finding the maximum *DD* value in (1) is equivalent to finding the minimum *NLDD* value in (2).

The Quasi-Newton algorithm [13], a two-step gradient descent search, is able to find the point with the minimum *NLDD* value in (2). We start the search from every instance in all positive bags with the same initial weights to find its corresponding local minimum and associated weights. Unlike the DD and EM-DD methods, where the global minimum of all local minima represents the region of interests, our MIL-based method uses all the distinct local minima, called IPs, to create the image bag features. These IPs are selected from all local minima by the following two conditions: (1) they need to be far away from each other in the bag feature space; and (2) they need to have small *NLDD* values. Given IPs = $\left\{(x_k^*, w_k^*) : k = 1, \cdots, m\right\}$ where x_k^*'s are the IP's feature values, w_k^*'s are the IP's feature weights, and *m* is the number of IPs, the bag feature $\phi(B_i)$ of image *i* (i.e., bag B_i) is calculated as:

$$\phi(B_i) = \begin{bmatrix} \min_{j=1,\cdots,N_i} \left\| x_{ij} - x_1^* \right\|_{w_1^*} \\ \vdots \\ \min_{j=1,\cdots,N_i} \left\| x_{ij} - x_m^* \right\|_{w_m^*} \end{bmatrix} \tag{3}$$

where x_{ij} is the j^{th} regional color and texture features of image *i*, N_i is the number of segmented regions in image *i*, and $\| \|_{w^*}$ represents the weighted Euclidean distance.

Support Vector Machines (SVMs). SVMs have been successfully used in many applications and are adopted in our proposed system. A set of SVMs are used to train the bag features of all training images to find optimum hyperplanes, each of which separates training images in one category with the other categories 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 the following optimization problem:

$$\min_{\omega,b,\xi} (\frac{1}{2}\omega^T \omega + C\sum_{i=1}^{l}\xi_i), \quad \text{Subject to} \quad y_i(\omega^T \phi(x_i) + b) > 1 - \xi_i, \quad \xi_i > 0 \tag{4}$$

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 Gaussian radial basis function kernel are used in our system since they yield excellent results compared to linear and polynomial kernels [14].

Since the SVMs are designed for the binary classification, an appropriate multi-class method is needed to handle several classes as in image categorization. We use "one against the others" as it achieves comparable performance with a faster speed than "one against one". We further map the SVM outputs into probabilities [15] so that our system returns the likelihood of each category that an image may belong to.

2.2 Global-Feature-Based SVMs

Inaccurate image segmentation may make the MIL-based bag feature representation imprecise and therefore decrease the categorization accuracy. We add global-feature-based SVMs to address this problem. In order to compensate the limitations associated with the specific color space and the specific texture representation, we construct the global features in a different manner as used in creating the regional features. To this end, two MPEG-7 descriptors are adopted in our system.

The SCD is one of the four MPEG-7 normative color descriptors [16]. 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 64-bin SCD in our system.

The EHD is one of the three normative texture descriptors used in MPEG-7 [16], where 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 subimages. Based on the EHD, we construct gEHD (global EHD) and sEHD (semi-global EHD) to address the 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 subimages into 13 different clusters [16] and construct the EHD for each cluster. So the length of our modified EHD is 70 and the total length of our global feature is 134.

After the global features of all the training images are obtained, they are fed into another set of SVMs to find optimum hyperplanes to distinguish one category from the others. This set of SVMs is designed by the same approaches used in the MIL-based SVMs.

2.3 Fusion Approach

For each test image, two sets of image features (i.e., MIL-based features and global features) are generated and sent to two respective sets of SVMs. Let y_1 and y_2 respectively be the output vectors from the MIL-based and global-feature-based SVMs for a given test image. The final output vector y is obtained by:

$$y = w * y_1 + (1 - w) * y_2 \tag{5}$$

where w determines the contribution from the MIL-based SVMs and is empirically set to be 0.5 as shown in Section 3.2. Once the integrated decision values are obtained, they are mapped to the probability values by the method introduced in [15].

3 Experimental Results

To date, we have tested our categorization algorithm on 2000 general-purpose images from COREL database. These images have 20 distinct categories with 100 images in each category. These categories contain different semantics including Africa, beach, buildings, buses, dinosaurs, elephants, flowers, horses, mountains, food, dogs, lizards, fashion, sunsets, cars, waterfalls, antiques, battle ships, skiing, and deserts.

3.1 Categorization Results

To measure the effectiveness of the proposed system, we randomly choose 50 images from each category as training images and the remaining 50 images are used as the testing images. We repeat the above procedure 5 times and calculate the average categorization accuracy for each category.

The proposed system is compared with DD-SVM [10] and our implemented HistSVM [3]. For the first 10 categories, the overall average categorization accuracy of HistSVM, DD-SVM, and our systems over 5 runs is 79.8%, 81.5%, and 88.2%, respectively. Our system performs 10.5% better than the HistSVM system in terms of the overall accuracy. In addition, the feature length of HistSVM system is 4096, which is about 20 times longer than ours. Our system also improves the accuracy by 8.2% over the DD-SVM system, which is 9 times slower than our system.

Fig. 1 plots the average categorization accuracy for each predefined image category of our proposed system, DD-SVM system, and HistSVM system. It clearly illustrates that the proposed system achieves the best average accuracy in most categories.

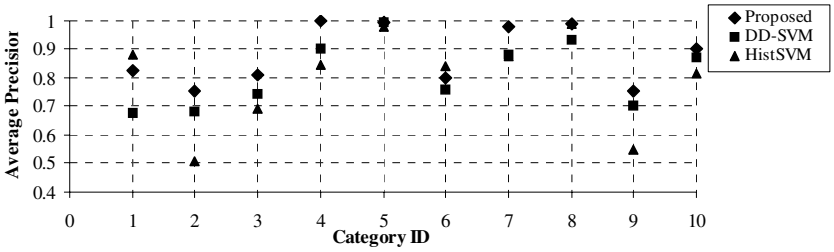

Fig. 1. Average categorization accuracy for each category by using three different methods

3.2 Validation of the Proposed Method

To verify the effectiveness of the proposed approach, the overall average categorization accuracy obtained by assigning different weights to the global-feature-based SVMs and the MIL-based SVMs is shown in Fig. 2, where G and R represent global and regional weight respectively. It clearly shows that our method (G:M = 0.5:0.5) achieves the best performance.

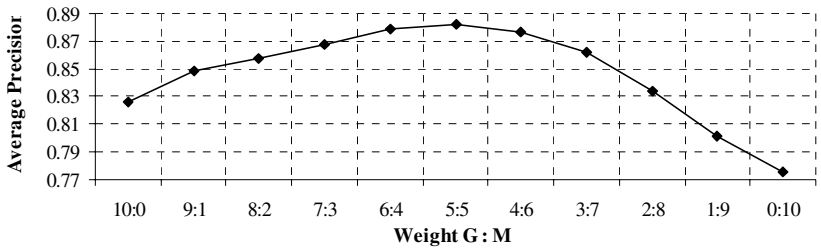

Fig. 2. Average categorization accuracy for different global and regional weights

It is observed that global-feature-based SVMs (i.e., G:R = 10:0) and MIL-based SVMs (i.e., G:R = 0:10) alone achieve the average accuracy of 82.6% and 77.6%, respectively. It clearly shows the effectiveness of the fusion approach as it improves the global and regional SVMs by 6.8% and 13.7% respectively. In addition, our global-feature-based SVMs system alone achieves better accuracy than both DD-SVM and HistSVM systems.

3.3 Sensitivity to the Number of Categories

The scalability of the method is tested by performing image categorization experiments over data sets with different numbers of categories. A total of 11 data sets are used in the experiments. The number of categories in a data set varies from 10 to 20. These data sets are arranged in the same manner as in [11] for fair comparisons. That is, the first 10 categories form the first data set; the first 11 categories form the second data set; etc. The average classification accuracy of our system and DD-SVM system by running 5 times on each of the 11 data sets is shown in Figure 3.

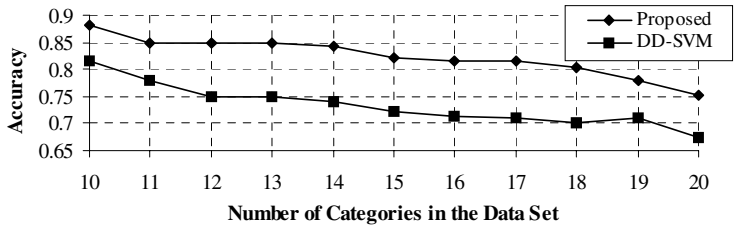

Fig. 3. Comparison of the two methods on the robustness to the number of categories

We observe a decrease in average categorization accuracy as the number of categories increases. When the number of categories becomes doubled (increasing from 10 to 20 categories), the average categorization accuracy of our proposed system and the DD-SVM system drops respectively from 88.2% to 75.3% and from 81.5% to 67.5%. However, our method outperforms DD-SVM consistently.

4 Conclusions

In this paper, we present an efficient and effective automatic image categorization system, which integrates MIL-based SVMs with global-feature-based SVMs. The main contributions are:

- EM-DD algorithm is used to find IP (Instance Prototypes) and the IP-based image bag features are further combined with SVMs to partly solve the problem of inaccurate image segmentation.
- Global-feature-based SVMs are integrated with MIL-based SVMs to further address the issues associated with inaccurate image segmentation, where global features are different from the regional features so that the limitations associated with specific color space and specific texture representation are also addressed.
- Multi-category SVMs are used to classify images by a set of confidence values for each possible category.

The proposed system has been validated by testing with 2000 general-purpose images with 20 distinct categories. The experimental results indicate that our system outperforms peer systems in the literature in terms of both accuracy and efficiency.

The proposed system can be easily integrated into the image retrieval system, where both 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 that categorization accuracy can be further improved.

References

1. Huang, J., Kumar, S., Zabih, R.: An automatic hierarchical image classification scheme. Proc. of 6th ACM Int'l Conf. on Multimedia (1998) 219-228
2. Chapelle, O., Haffner, P., Vapnik, V.: Support vector machines for histogram-based image classification. IEEE Trans. on Neural Networks 10(1999) 1055–1064
3. Smith, J. R., Li, C. S.: Image classification and querying using composite region templates. Int'l J. Computer Vision and Image Understanding 75(1999) 165-174
4. Barnard, K., Forsyth, D.: Learning the semantics of words and pictures. Proc. Int'l Conf. Computer Vision 2(2001) 408-415
5. Jeon, J., Lavrenko, V., Manmatha, R.: Automatic image annotation and retrieval using cross-media relevance models. Proc. 26th Intl. ACM SIGIR Conf. (2003) 119–126
6. Li, J., Wang, J. Z.: Automatic linguistic indexing of pictures by a statistical modeling approach. IEEE Trans. on PAMI 25(2003) 1075-1088
7. Murphy, K, Torralba, A., Freeman, W.: Using the forest to see the trees: a graphical model relating features, objects, and scenes. Advances in Neural Information Processing Systems, Vol. 16, Cambridge, MA, MIT Press (2004)

8. Maron O., Ratan, A. L.: Multiple-instance learning for natural scene classification. Proc. 15th Int'l Conf. Machine Learning (1998) 341-249
9. Zhang, Q., Goldman, S. A., Yu, W., Fritts, J.: Content-based image retrieval using multiple instance learning. Proc. 19th Int'l Conf. Machine Learning (2002) 682-689
10. Andrews, S., Tsochantaridis, I., Hofmann, T.: Support vector machines for multiple-instance learning. Advances in Neural Information Processing Systems 15, Cambridge, MA: MIT Press (2003)
11. Chen, Y., Wang, J. Z.: Image categorization by learning and reasoning with regions. Journal of Machine Learning Research 5(2004) 913-939
12. Dietterich, T. G., Lathrop, R. H. , Lozano-Perez, T., Solving the multiple-instance problem with axis-parallel rectangles. Artificial Intelligence 89(1997) 31-71
13. Press, S. A., Teukolsky, S. A., Vetterling, W. T., Flannery, B. P.: Numerical recipes in C: the art of scientific computing. Cambridge Univeristy Press, New York (1992)
14. Scholkopf, B., Sung, K., Burges, C., Girosi, F., Niyogi, P., Poggio, T., Vapnik, V., Comparing Support Vector Machines with Gaussian Kernels to Radial Basis Function Classifiers. MIT, A.I. Memo 1599 (1996)
15. Platt, J. C.: Probabilistic Output for Support Vector Machines and Comparisons to Regularized Likelihood Methods. In Bartlett, A., Schölkopf, P., Schuurmans, B. (eds.): Advances in Large Margin Classifiers. MIT Press Cambridge, MA (2000)
16. Manjunath, B. S., Salembier, P., Sikora, T.: Introduction to MPEG-7 Multimedia Content Description Interface. John Wiley & Sons (2002)