# A COMPLEMENTARY SVMS-BASED IMAGE ANNOTATION SYSTEM

*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**

A novel automatic image annotation system is proposed, which integrates two sets of SVMs (Support Vector Machines), namely the MIL-based (Multiple Instance Learning) and global-feature-based SVMs, for annotation. The MIL-based features are obtained by applying MIL on the image blocks. They are further input to a set of SVMs for finding the optimum hyperplanes to categorize training images. Similarly, global color and texture features are fed into another set of SVMs for categorizing training images. Consequently, two sets of image features are constructed for each test image and are respectively sent to the two sets of SVMs, whose outputs are incorporated to obtain the final annotation results. Our system is validated using COREL images and outperforms the peer systems in terms of efficiency and accuracy.

## 1. INTRODUCTION

The term image annotation refers to the labeling of images with relevant keywords. It is very important for image retrieval and object recognition. Since manual annotation is expensive and subjective, automatic image annotation becomes the research focus of a number of researchers. A few relevant systems are briefly reviewed here.

Chapelle *et al.* [1] apply SVMs on the global 16×16×16-bin HSV color histograms to annotate images. The ALIP system [2] uses the 2D MHMM (Multi-resolution Hidden Markov Model) on color and texture features of image blocks of size 4×4 for automatic image annotation. More recently, the MIL technique has been applied for automatic image annotation. Maron and Ratan [3] use the DD (Diverse Density) learning algorithm for natural scene classification. Zhang and Goldman [4] develop the EM-DD algorithm by extending the DD algorithm with EM (Expectation Maximization) so that the algorithm can scale up to large data sets and run faster. In both DD and EM-DD algorithms, the goal is to find an optimum point in the feature space with a global maximum DD value to represent the object of interest. However, this optimum point may not correctly represent the object of interest due to the imperfect or incorrect image segmentation. To address this issue, Chen and Wang [5] propose the DD-SVM method, which combines EM-DD with SVMs, to construct bag-based image features using multiple local maxima instead of one global maximum. These features are fed into SVMs to form the hyperplanes for image classification and annotation.

In spite of their successes, all these annotation systems have shortcomings. Global-feature based systems cannot precisely represent the objects, which correspond to the semantics of an image. Block-based systems often break an object into several blocks or put different objects into a single block. Region-based systems have the similar problems due to inaccurate image segmentation.

In this paper, we propose a novel fusion approach which combines MIL-based and global-feature-based SVMs for effective and efficient image annotation. Instead of segmenting an image into homogeneous regions, the image is divided into several blocks. MIL is applied to the image blocks to obtain bag features, which are input to a set of SVMs for finding the optimum hyperplanes to categorize training images. To address the possible RST (rotation, scaling, and translation) variant issues, we create the global-feature-based SVMs for categorizing training images, where global color and edge histograms are the inputs. For each test image, two sets of image features are constructed and sent to the respective sets of SVMs. The outputs from these SVMs are incorporated to obtain the final annotation 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

*2.1.1. Image sub-blocking and block feature extraction*

It is well known that image segmentation is a difficult task and no system can achieve perfect segmentation. In

0-7803-9134-9/05/$20.00 ©2005 IEEE

addition, it is computationally expensive. As a result, image segmentation is not performed on the proposed system. Instead, we evenly divide the image into 9 non-overlapping blocks as shown in Fig.1. Several layouts of the image blocks are further considered in our system based on the observation that the main object usually locates around the center. 1) Center block 5 is respectively shifted to the left, right, up, and down by a half block to accommodate the possible shifts of the main object. 2) Three horizontal groupings ({1,2,3}, {4,5,6}, {7,8,9}) and three vertical groupings ({1,4,7}, {2,5,8}, {3,6,9}) are considered for the possible shifts of a wide or tall object. Therefore, each image is represented by 19 blocks. For each block, the mean and standard deviation in LUV color space are computed as the color feature. The texture feature is calculated by the average energy in each high frequency band after the 2-level wavelet decomposition.

| | | |
|---|---|---|
| 1 | 2 | 3 |
| 4 | 5 | 6 |
| 7 | 8 | 9 |

Fig. 1: Image sub-blocking

*2.1.2. Multiple-instance learning (MIL) and bag features*

In MIL, the user labels the bag (i.e., image), which usually contains many instances (i.e., regions), as positive or negative. The goal of the MIL is to find what is common in all positive images, but not in any negative images. The DD method [3] converts this goal to a maximization problem. That is, suppose we have $n$ labeled bags and the hypothesis $t$, the DD value is:

$$DD(t) = \prod_{i=1}^{n} \Pr(B_i, l_i \mid t) = \prod_{i=1}^{n} (1 - |l_i - Label(B_i \mid t)|)$$
$$Label(B_i \mid t) = \max_j \left\{ \exp[-\sum_{d=1}^{m} (s_d (B_{ijd} - t_d))^2] \right\} \quad (1)$$

where $B_i$ refers to the $i^{th}$ bag, $l_i$ refers to the actual label (0 or 1) of the $i^{th}$ bag, $B_{ij}$ refers to the $j^{th}$ instance of bag $i$, $S_d$ refers to the feature weight $S$ on dimension $d$ as different features may have different weights, and $t_d$ is the value of $t$ on dimension $d$. The maximization of (1) is to find the optimum $t$ that leads to the maximum DD value. It is easily observed by analyzing (1) that if any instance in the negative bag $B_i$ is close to $t$, the $Label(B_i \mid t)$ will be close to 1 and the $\Pr(B_i, l_i \mid t)$ (i.e., the probability of the image being negative under the hypothesis $t$) will be close to 0. This will adversely drop the DD value close to 0 even if all the other negative instances are far away from $t$. As a result, we modify (1) so the DD value will not be dramatically affected by several abnormal instances as explained above. The new definition of DD is:

$$DD(t) = \sum_{i=1}^{n} \Pr(B_i, l_i \mid t) \quad (2)$$

A simplex search method [6] is applied on (2) to find the optimum point $t$ which yields the maximum DD value. This method is faster than the gradient-based method as used in other annotation systems [3-5] with a comparable accuracy since it is a direct search without using any numerical or analytical gradients.

For finding each local maximum and its corresponding weight, we start the search from every instance of all positive bags with the same initial weights. Once all local maxima are found, IPs (Instance Prototypes) are obtained by replacing clumped local maxima with their average and removing local maxima whose DD values are too small. These IPs $\{(x_k^*, w_k^*) : k = 1, \cdots n\}$ approximately represent all possible objects of interest and are further used to construct the bag feature $\phi(B_i)$ for image $B_i = \{x_{ij} : j = 1, \cdots N_i\}$:

$$\phi(B_i) = \begin{bmatrix} \exp\left( \min_{j=1,\cdots,N_i} \left\| x_{ij} - x_1^* \right\|_{w_1^*} / 30 \right) \\ \vdots \\ \exp\left( \min_{j=1,\cdots,N_i} \left\| x_{ij} - x_n^* \right\|_{w_n^*} / 30 \right) \end{bmatrix} \quad (3)$$

where $x_k^*$'s are the IP's feature values, $w_k^*$'s are the IP's feature weights, $n$ is the number of IPs, $x_{ij}$ is the $j^{\text{th}}$ block features of image $i$, $N_i$ is the block numbers (i.e., 19) in image $i$, and $\| \|_{w^*}$ is the weighted Euclidean distance. The exponential function is used here to properly scale the values of the bag features to the range between 0 and 1.

*2.1.3. Support vector machines (SVMs)*

The bag features for all training images are obtained using (3) and are further fed into SVMs, which will find a hyperplane that separates the training data by a maximal margin. That is, given $m$ training data $\{x_i, y_i\}$'s, where $x_i \in R^n$ and $y_i \in \{-1, 1\}$, SVMs need to solve the following optimization problem:

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

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 non-linear SVMs with the Gaussian radial basis function kernel are used in our system since they achieve excellent results compared to linear and polynomial kernels [7].

A fast multi-class method, namely "one against the others" (i.e., compare a given class with all the others put together), is used in our system to handle several classes as in image annotation. We also map the SVM outputs into probabilities [8] to reflect the likelihood of each category that an image may belong to.

**2.2. Global-feature-based SVMs**

Chen and Wang [5] negate the labels of all bags and repeat the gradient-based search to improve the annotation

accuracy by around 2.2% for the 10-class case. However, the training will be at least 9 times slower when compared with the scheme without negation. In our system, we add global-feature-based SVMs, which require almost no additional training time, to get much better improvement. These SVMs can also compensate the RST variant issues associated with the block scheme.

Our global features combine MPEG-7 SCD (Scalable Color Descriptor) and modified EHD (Edge Histogram Descriptor) to compensate the limitations related to the specific color and texture representations.

The SCD is one of the four MPEG-7 normative color descriptors [9]. It uses the HSV color histograms to represent an image since the HSV color space approximates human's perception. We directly adopt the 32-bin SCD in our system. The EHD is one of the three normative texture descriptors used in MPEG-7 [9]. It captures the spatial distribution of edges in an image. Five types of edges (i.e., 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 the 5-bin global EHD to represent the edge distribution of the entire image. Therefore, the total length of our global feature is very compact (i.e., 32+5=37).

The global features of all the training images are also fed into another set of SVMs. The same multi-class and mapping methods are used to obtain the likelihood of each category that an image may belong to.

### 2.3. Fusion approach

The fusion approach combines the outputs from the MIL-based and global-feature-based SVMs to obtain the final annotation results. For each test image, two sets of image features are constructed and sent to the corresponding SVMs. The scores from the MIL-based SVMs ($y_1$) and the global-feature-based SVMs ($y_2$) are combined to obtain the final score $y$ as follows:

$$y = w*y_1 + (1-w)*y_2 \qquad (5)$$

where $w$ determines the contribution from the MIL-based SVMs and is empirically set to be 0.6. Once the fusion result is obtained, it is mapped to probability output according to the method mentioned in [8].

## 3. EXPERIMENTAL RESULTS

We have tested our annotation 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 people and villages, beach, historical buildings, buses, dinosaurs, elephants, flowers, horses, mountains, food, dogs, lizards, fashion, sunsets, cars, waterfalls, antiques, battle ships, skiing, and deserts.

### 3.1. Annotation results

To measure the effectiveness of our system, we randomly choose 50 images from each category as training images and the rest are used as test images. We repeat the above procedure 5 times and calculate the average annotation accuracy. Table 1 shows the average annotation results based on images from the first 10 categories.

The proposed system is also compared with DD-SVM [5] and our implemented HistSVM [1] using images from the first 10 categories. The overall average annotation accuracy of HistSVM, DD-SVM and our systems over 5 runs is 79.8%, 81.5%, and 88.1%, respectively. Our system performs better than the HistSVM system with an 8.3% difference in the overall accuracy. In addition, the feature length of our system is around 37+93=130, which is about 30 times shorter than the ones from HistSVM. Our system also improves the accuracy by 8.2% over the DD-SVM system, which is about 9 times slower than our system. In addition, the average elapse time for one query of our system is about 35ms, which is very suitable for online processing. Fig. 2 plots the average annotation accuracy for each of the first 10 predefined categories by using the above three systems. It clearly illustrates that the proposed system achieves the best average accuracy in all categories except category 1, where HistSVM achieves the best accuracy. This is probably because most images in category 1 have distinct colors.

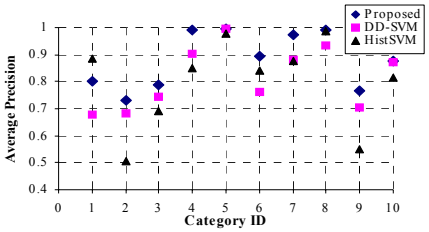


Fig. 2: Comparisons of the average annotation accuracy for each category by using 3 methods

To verify the effectiveness of the fusion approach, the overall average annotation accuracy obtained by assigning different weights to the global-feature-based and MIL-based SVMs is plotted in Fig. 3, where G and M represent global and MIL weights, respectively. It clearly shows that our method (G:M = 4:6) performs the best, and the global-feature-based SVMs (G:M = 10:0) and the MIL-based SVMs (G:M = 0:10) alone achieve the average accuracy of 81.4% and 84.3%, respectively.

### 3.2. Sensitivity to the number of categories

Table 1: Confusion matrix of the proposed system, where each row lists the average percentage of the images in one category annotated into each of the 10 categories. Numbers on the diagonal show the annotation accuracy for each category.

| | Africa | Beach | Build. | Buses | Dino. | Eleph. | Flower | Horse | Mount. | Food |
|---|---|---|---|---|---|---|---|---|---|---|
| Africa | **0.8** | 0.012 | 0.056 | 0.028 | 0.012 | 0.048 | 0.004 | 0.008 | 0.004 | 0.028 |
| Beach | 0.02 | **0.732** | 0.06 | 0.028 | 0 | 0.024 | 0.012 | 0.008 | 0.112 | 0.004 |
| Build. | 0.056 | 0.036 | **0.788** | 0.024 | 0.004 | 0.036 | 0.004 | 0 | 0.032 | 0.02 |
| Buses | 0 | 0 | 0 | **0.992** | 0 | 0 | 0 | 0 | 0 | 0.008 |
| Dino. | 0 | 0 | 0 | 0 | **0.996** | 0.004 | 0 | 0 | 0 | 0 |
| Eleph. | 0.004 | 0.004 | 0.02 | 0 | 0 | **0.896** | 0 | 0.024 | 0.02 | 0.032 |
| Flower | 0.008 | 0 | 0 | 0 | 0 | 0 | **0.972** | 0.008 | 0.004 | 0.008 |
| Horse | 0 | 0.008 | 0 | 0 | 0 | 0 | 0 | **0.992** | 0 | 0 |
| Mount. | 0 | 0.148 | 0.02 | 0.012 | 0 | 0.036 | 0.008 | 0.004 | **0.764** | 0.008 |
| Food | 0.04 | 0.008 | 0 | 0.016 | 0.024 | 0 | 0.004 | 0.008 | 0.024 | **0.876** |

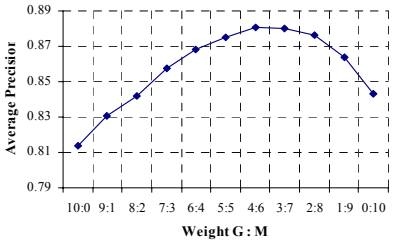


Fig. 3: Average annotation accuracy for different weights

The scalability of the method is tested by performing image annotation 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 [5] 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 annotation accuracy of our system and the DD-SVM system by running 5 times on each data set is shown in Fig. 4. We observe a decrease in average annotation accuracy in both systems as the number of categories increases. However, our method consistently outperforms DD-SVM in all data sets.

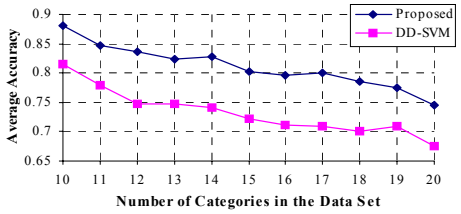


Fig. 4: Comparison of the two methods on the robustness to the number of categories

## 4. CONCLUSIONS

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

- Novel block-based features, instead of the expensive segmentation-based features, are used for MIL.
- More robust DD definition is employed in MIL.
- A faster search algorithm (i.e., a simplex method) is applied to speed up the training procedure.
- The IPs-based bag features are combined with SVMs to approximately represent all possible objects of interest.
- The global-feature-based SVMs are integrated with the MIL-based SVMs to address the RST related issues.
- The global and block features are constructed in a different manner to compensate the limitations of the specific color and texture representations.

The proposed system can be easily integrated into the image retrieval system, where both annotated keywords and the query image(s) can be combined as the query.

## 5. REFERENCES

[1] O. Chapelle, P. Haffner, and V. N. Vapnik, "Support Vector Machines for Histogram-Based Image Classification," IEEE Trans. Neural Networks, vol. 10, pp. 1055–1064, 1999.

[2] J. Li and J. Z. Wang, "Automatic Linguistic Indexing of Pictures by a Statistical Modeling Approach," IEEE Trans. PAMI, vol. 25, pp. 1075-1088, 2003.

[3] O. Maron and A. L. Ratan, "Multiple-Instance Learning for Natural Scene Classification," Proc. of 15th Int'l Conf. Machine Learning, pp. 341-249, 1998.

[4] Q. Zhang, S. A. Goldman, W. Yu, and J. Fritts, "Content-Based Image Retrieval Using Multiple Instance Learning," Proc. of the 19th Int'l Conf. Machine Learning, pp. 682-689, 2002.

[5] Y. Chen and J. Z. Wang, "Image Categorization by Learning and Reasoning with Regions," J. of Machine Learning Research, vol. 5, pp. 913-939, 2004.

[6] J. C. Lagarias, J. A. Reeds, M. H. Wright, and P. E. Wright, "Convergence Properties of the Nelder-Mead Simplex Method in Low Dimensions," SIAM J. of Optimization, vol. 9, pp.112-147, 1998.

[7] 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," MIT, A.I. Memo 1599, 1996.

[8] 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.

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