Region-based Image Annotation using Asymmetrical Support Vector Machine-based Multiple-Instance Learning

Changbo Yang and Ming Dong
Machine Vision and Pattern Recognition Lab.
Department of Computer Science
Wayne State University
Detroit, MI 48202

Jing Hua
Department of Computer Science
Wayne State University
Detroit, MI 48202

Abstract

In region-based image annotation, keywords are usually associated with images instead of individual regions in the training data set. This poses a major challenge for any learning strategy. In this paper, we formulate image annotation as a supervised learning problem under Multiple-Instance Learning (MIL) framework. We present a novel Asymmetrical Support Vector Machine-based MIL algorithm (ASVM-MIL), which extends the conventional Support Vector Machine (SVM) to the MIL setting by introducing asymmetrical loss functions for false positives and false negatives. The proposed ASVM-MIL algorithm is evaluated on both image annotation data sets and the benchmark MUSK data sets.

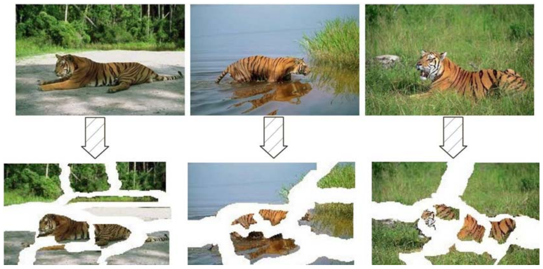


Figure 1. Three sample images of "tiger" (top row) and their segmented regions (bottom row). A large number of irrelevant noisy regions, such as "grass", "water", and "bush", exists in the training images for the keyword "tiger".

1. Introduction

With the rapid development of digital photography, large collections of digital pictures have sprung up easily in recent years and users would like to browse and search these collections. Consequently, Content-Based Image Retrieval (CBIR) has attracted significant interest amongst the computer vision community. CBIR systems use low-level features automatically extracted from images/image regions, such as color and texture, to search for images relevant to a query. However, CBIR systems often require users to pose image queries using low-level features, which is difficult for most people to do.

An ideal image retrieval system from a user perspective would involve what is referred as semantic retrieval, where the user makes a request like "find pictures of the sky" instead of "find pictures of predominantly blue and white". The traditional "low-tech" solution to semantic retrieval is to annotate each image manually with keywords and then search on those keywords using a text search engine. The underlying principle of this approach is that keywords can capture the semantic content of images more precisely, and thus provide better means to organize and search an image database. However, manual annotation is not scalable and very expensive when the volume of data becomes very large. An automated reverse process that discovers the "words" associated with a picture by human viewers is highly desired to handle the massive digital image resources. Automatic image annotation is a process in which a computer program learns the relationship between the content of an image and its semantic meaning, and assigns keywords to the image accordingly. An image contains several regions. Since each region may have different contents and represent different semantic meaning, it is intuitive to divide an image into regions and extract visual features from each region. This is usually the first step of region-based image annotation. A statistical model is then learnt from a set of annotated training images to link image regions to keywords and produce the annotation for a testing image. However, a major hurdle remains in the aforemen-

Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)
0-7695-2597-0/06 $20.00 © 2006 IEEE

tioned learning infrastructure. With few exceptions, the annotation information for a training image is available only at the *image level*, but NOT at the *region level* [9, 16]. In other words, keywords are associated with images instead of individual regions. For example, the top row of Figure 1 shows four images of "tiger" and the bottom row shows the corresponding image regions segmented using Normalized-cuts [14]. The existence of a large number of irrelevant regions in the training data, such as "grass", "water", and "bush", poses a major challenge for any statistical learning model. To find the correct correspondence between an image region and the keyword "tiger", a learner must be able to differentiate "tiger" regions from other noisy regions at the outset.

In this paper, we formulate image annotation as a supervised learning problem under Multiple-Instance Learning (MIL) framework, and present a novel Asymmetrical Support Vector Machine-based MIL algorithm (ASVM-MIL), which extends the conventional Support Vector Machine (SVM) to the MIL setting through the introduction of asymmetrical loss functions for false positive and false negative examples. By maximizing the pattern margins subject to the MIL constraints, ASVM-MIL converts the MIL problem to a traditional supervised learning problem, and thus can take the advantage of strong learning capability of SVM.

The rest of this paper is organized as follows. Section 2 reviews the related work. In Section 3, we formulate image annotation as a supervised learning problem in the MIL setting. ASVM-MIL algorithm are presented in Section 4. Section 5 describes the extensive experiments we have performed and provides the results. Finally, we conclude in Section 6.

2. Related Work

In this section, we provide a review of previous works in region-based image annotation and MIL.

2.1. Region-based Image Annotation

Region-based image annotation is a highly challenging problem because of the semantic gap between low-level image contents and high-level concepts. Starting from a training set of annotated images, many statistical learning models have been proposed in the literature to associate region-based visual features with semantic concepts (keywords). Most of the recent efforts followed an unsupervised learning approach. The key idea is to run a clustering algorithm on the low-level feature space, and then estimate the joint density of keywords and low-level features [4].

Specifically, Mori et al. [13] used a Co-occurrence Model in which they looked at the co-occurrence of words with image regions created using a regular grid. Duygulu et al. [8] proposed to describe images using a vocabulary of blobs. Their Translation Model assumes that image annotation can be viewed as a task of translating a vocabulary of blobs to a vocabulary of words. Barnard et al. [2] extended the machine translation method through a hierarchical clustering model and developed several models for the estimation of joint distribution between a region and a keyword. Jeon et al. [11] introduced a cross-media relevance model (CMRM) that learns the joint distribution of a set of regions (blobs) and a set of keywords rather than the correspondence between a single region (blob) and a single keyword.The CMRM modelling was subsequently improved through a continuous-space relevance model and a multiple Bernoulli relevance model. Blei et al. [3]proposed three hierarchical probabilistic mixture models for image annotation, in which the joint probability between words and regions are modelled by different latent variables.

Since these unsupervised approaches rely on clustering as the basis for image annotation, the performance of annotation is strongly influenced by the quality of unsupervised learning. Currently, most approaches perform region clustering based on visual features and suffer from the semantic gap, i.e., regions with different semantic concepts but similar appearance may be grouped together, while regions with the same semantic content may be separated into different clusters due to diverse appearance. To circumvent this difficulty, more recently, region-based annotation problem has been formulated as a supervised learning problem by imposing strict semantic constraints on the training data [4, 17].

2.2. Multiple-Instance Learning

MIL is a variation of supervised learning [12], where the task is to learn a concept given positive and negative bags of instances. It stems from the pioneering paper by Dietterich et al., which introduced the Axis-Parallel hyper-Rectangle (APR) algorithm [7] and the MUSK data sets. As a parametric approach, the objective of APR algorithm is to find the parameters that, together with the MI assumptions, can best explain the training data. Following this work, there is a significant amount of research directed towards the extensions of the APR algorithm.

The first probability model of MIL is the Diverse Density (DD) model proposed in [12]. The idea of the method is to examine the distribution of instances, and look for a point in the feature space that is close to all instances in the positive bags and far from those of negative bags.

The new trend of MIL is to upgrade single-instance learning methods to deal with MIL problems, such as decision trees , nearest neighbor, neural networks, and SVMs [1]. In particular, Andrews et al. proposed two SVM-based formulations of MIL, mi-SVM and MI-SVM [1]. To solve the maximum margin problem under the MIL constraints, both algorithms modify the conventional SVM through a iterative heuristic optimization. More recently, Chen and Wang proposed a DD-SVM algorithm [6]. DD-SVM assumes that the classification of the bags is only related to some properties of the bags. Consequently, it solves the MIL problem by transforming the original feature space to a new bag feature space, and training a SVM in the new space.

3. Region-based Image Annotation and MIL

In this section, we will formulate image annotation as a supervised learning problem in the MIL setting.

3.1. Image Partition

An image usually contains several regions. Since regions may have different contents and represent different semantic meaning, a straightforward solution is dividing an image into regions and extracting visual features from each region. The image regions could be determined through either image segmentation or image cutting. For example, Duygulu et al. [8] used Normalized-cuts [14] to obtain image regions. For each segmented region, features such as color, texture, position, and shape are computed. On the other hand, Feng et al. [10] cut an image into many grids and treat each grid as a region.

3.2. Image Annotation as a Supervised Learning Problem

Let J denote the testing set of un-annotated images, and let T denote the training collection of annotated images. Each testing image $I \in J$ is represented by its regional visual features $r = \{r_1 \ldots r_m\}$, and each training image $I \in T$ is represented by both a set of regional visual features $r = \{r_1 \ldots r_m\}$ and a keyword list $W_I \subseteq V$, where r_j $(j = 1 \ldots m)$ is the visual features for region j, $V = \{w_1 \ldots w_n\}$ the vocabulary, and w_i $(i = 1 \ldots n)$ the ith keyword in V.

The goal of image annotation is to select a set of keywords W that best describes a given image I from the vocabulary V. The training set, T, consists of N image-keyword pairs $T = \{(I_1, W_1), ..., (I_N, W_N)\}$. If we treat each keyword w_i as a distinct class label, the annotation problem can be converted to an image classification problem and stated as follows: *Given the feature vector of a testing image I, which class (keyword)*[1] w_i *does I belong to?*

However, we should notice that the image annotation is not a traditional supervised learning problem because the training image set does not provide explicit correspondence between keywords and regions. Keywords are associated with images instead of individual regions, which presents a major hurdle for both approaches. As shown in Figure 1, the images annotated with keyword "tiger" may contain many other regions that correspond to keywords "grass", "river", or "bush". The low-level features in these irrelevant regions may be completely different from "tiger" regions. The large amount of noise existing in the training data will also present a major difficulty for a non-parametric classifier, such as SVM. To circumvent this problem, we formulate region-based image annotation in the MIL setting in the next Section.

3.3. MIL and Image Annotation

In the MIL setting, each bag may contain many instances, but a bag is labelled positive as long as one of its instances is positive. A bag is labelled negative only if its instances are all negatives. From a collection of labelled bags, the learner tries to induce a concept that will label individual instances correctly. This problem is even harder than noisy supervised learning since the ratio of negative to positive instances in a positive bag (the noise ratio) can be arbitrarily high.

In region-based image annotation, each region is an *instance*, and the set of regions that comes from the same image can be treated as a *bag*. We annotate an image by keyword w_i if at least one region in the image has the semantic meaning of w_i. For example, the first image in Figure 1 is annotated with keyword "tiger" and segmented to ten regions. These ten regions consist of a positive bag for "tiger". In this positive bag, there are only two positive instances because only two regions are actually relevant to "tiger". Given an image labelled by keyword w_i, we can expect that at least one region will correspond to w_i even if segmentation may be imperfect. Hence, the image annotation problem is in essence identical to the MIL setting.

4. ASVM-MIL Algorithm

In this section, we present a ASVM-MIL algorithm for region-based image annotation. ASVM-MIL takes the advantage of the strong generalization capability of SVM to find a optimal nonlinear decision boundary for each keyword, and thus greatly facilitates image annotation.

4.1. ASVM-MIL

SVM is a well-accepted machine learning algorithm that tries to find a separating hyperplane (w, b) with maxi-

[1] In image annotation, I may be labelled by more than one keyword. Consequently, I may belong to several classes in the image classification problem.

mum margins between positive instances and negative instances. In the non-separable cases, the algorithm tries to find the decision boundary leading to the minimum classification errors. Assuming a binary classification problem based on independent identically distributed (i.i.d.) data $(x_1, y_1), \ldots, (x_l, y_l) \in X \times Y, Y = \{-1, 1\}$, the task for finding the optimal hyperplane is to minimize the following objective function,

$$\begin{aligned} &minimize < w.w > + C \sum_{i=1}^{l} \xi_i^2, && (1) \\ &s.t. \quad y_i(x_i w + b) \geq 1 - \xi_i; \quad \xi_i \geq 0 \quad y_i \in \{\pm 1\} \end{aligned}$$

where ξ_i is a slack parameter that allows classification errors. The dual optimization problem is given as,

$$\begin{aligned} \max_{\boldsymbol{\alpha}} W(\boldsymbol{\alpha}) = &\sum_{k=1}^{l} \alpha_k - \frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} \alpha_i \alpha_j y_i y_j < x_i, x_j > \\ &- \frac{1}{2C} < \boldsymbol{\alpha} \cdot \boldsymbol{\alpha} > \qquad (2) \end{aligned}$$

with constraints,

$$s.t. \quad \sum_{i=1}^{l} \alpha_i y_i = 0, \quad \alpha_i \geq 0, \quad i = 1, \ldots, l \qquad (3)$$

where α_i is the Lagrange multiplier. Even though SVM has been successfully applied to many classification problems, directly using SVM in MIL is not feasible due to the existence of noisy instances in positive bags. In [12], it was shown that the result of directly applying SVM in MIL is not as good as some MIL specified algorithms.

Actually, if we directly employ SVM in MIL, we have to minimize the classification error of bags. The corresponding SVM can be obtain by solving the following optimization problem,

$$\begin{aligned} &minimize < w.w > + C E_B, && (4) \\ &s.t. \quad y_i(x_i w + b) \geq 1 - \xi_i; \quad \xi_i \geq 0 \quad y_i \in \{\pm 1\} \end{aligned}$$

where E_B is the number of the bag label errors. To solve Equation 4, we have to propagate the errors from the bag level to the instance level. Unfortunately, instance level labels are not available for MIL problems so that Equation 4 leads to a mixed integer programming problem, which requires the maximum decision margin as well as the optimal instances labelling [1]. The resulting problem is a NP-complete problem, and can not be solved efficiently. Next, we present an approximated solution using ASVM.

E_B consists of two kinds of errors, false positive E^+ and false negative E^-. The objective of ASVM is to introduce different loss functions for false positives and false negatives, which translates into a bias for larger multipliers for the class where the cost of misclassification is higher. Let C_1 and C_2 denote the penalty for false positives and false negatives, respectively, Equation 4 is modified as,

$$\begin{aligned} &minimize < w.w > + C_1 E^+ + C_2 E^-, && (5) \\ &s.t. \quad y_i(x_i w + b) \geq 1 - \xi_i; \quad \xi_i \geq 0 \quad y_i \in \{\pm 1\} \end{aligned}$$

Since the number of the true positive instances in positive bags is unknown in the MIL setting, theoretically, it is impossible to determine the relative weighting between C_1 and C_2. However, it is generally reasonable to assume that the average number of positive instances in each positive bag is greater than one. So, a false negative does not necessarily give a bag label error, while a false positive will certainly lead to an error. Thus, C_1 should be greater than C_2. Without loss of generality, we can safely assume that C_1 takes a positive value and $C_2 = 0$. With this assumption and converting Equation 5 to the soft margin constraints, ASVM is obtained by solving the following optimization problem,

$$\begin{aligned} &minimize < w.w > + C \sum_{i=1}^{l} \xi_i^2, && (6) \\ &s.t. \quad y_i(x_i w + b) \geq 1 - \frac{(y_i + 1)}{2} \xi_i; \xi_i \geq 0 \quad y_i \in \{\pm 1\} \end{aligned}$$

where ξ_i is the slack variable and C is a constant that controls the tradeoff between the classification errors and the maximum margin. The primal formulation of the Lagrangian will be:

$$\begin{aligned} L_P = &\frac{||w||^2}{2} + C \sum \xi_i - \sum_{i=1}^{l} \alpha_i [y_i(w.x_i + b) - 1 \\ &+ \frac{(y_i + 1)}{2} \xi_i] \qquad (7) \end{aligned}$$

Using the derivatives $\frac{\partial Lp}{\partial w} = 0, \frac{\partial Lp}{\partial b} = 0, \frac{\partial Lp}{\partial \xi} = 0$ and the Kuhn-Tucker conditions, we get the dual formulation:

$$\begin{aligned} L_D = &\sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{l} \alpha_i \alpha_j y_i y_j [K(x_i, x_j) + \\ &\frac{(y_i + 1)(y_j + 1)}{4C} \delta_{ij}] \qquad (8) \end{aligned}$$

where δ_{ij} is the Kronecker δ defined as 1 if $i = j$ and 0 otherwise. Let $\theta = \frac{1}{C}$ in Equation 8, the difference between ASVM and SVM lies in the addition of $\theta \delta_{ij}$ to the kernel matrix. The new kernel matrix is given by:

$$K'(x_i, x_j) = K(x_i, x_j) + \theta \delta_{ij} \qquad (9)$$

Compared with other SVM-based MIL algorithms, ASVM-MIL has several advantages. For example, mi-SVM is known as the first algorithm that tries to solve MIL problem by a modified SVM. In mi-SVM, a SVM is trained in the instance feature space using all negative instances and selected positive instances. In addition, the algorithm uses a heuristic method to refine the decision boundary of SVM. The general scheme is to alternate the following two steps: 1) train a SVM for the given training data, and 2) once the discriminant is obtained, update the labels of one or several instances in the positive bags. Through this modification, the algorithm adjusts labels of the positive instances to make sure all positive bags follow the MIL setting. However, there is no guarantee that all negative instances will lie on the negative side. On the other hand, ASVM-MIL tries to minimize false positives (with higher misclassification cost) by directly modifying the margin constraints of SVM. In addition, ASVM-MIL does not require an iterative approach to find the final discriminant. Convergence is a not an issue with ASVM-MIL.

4.2. Parameter Estimation for ASVM

In ASVM, besides the kernel parameters for a regular SVM, θ also needs to be determined. The selection of the kernel parameters in SVM is a long-standing question. Empirically, cross-validation with grid-search is the most popular method [5]. ASVM introduces a new parameter θ to control the relative penalty between false positives and false negatives. For a technical point of view, we should select a θ that can best describe the training data in the MIL setting.

- Case 1: If each bag contains only one instance, the MIL problem will be reduced to a conventional supervised learning problem. Consequently, we can set $\theta = 0$ in the ASVM and the ASVM becomes a regular SVM. We denote corresponding θ as θ_0.
- Case 2: If each positive bag has a very large number of true positive instances, the loss caused by a false negative is much less than that caused by a false positive. In this case, the ideal decision boundary should be the one that produces no false positives and as few false negatives as possible. We denote corresponding θ as θ_1.

Generally speaking, the decision boundary will move towards the positive samples as θ increases, and consequently, the number of false positive will decrease while that of false negative will increase. For a given MIL problem, since the numbers of true positives in positive bags are unavailable, the optimal θ has to be selected empirically in the range $[\theta_0, \theta_1]$ on a separate validation set.

5. Experimental Results

In this section, we evaluate our MIL framework for image annotation based on a collection of images from COREL and the MUSK data sets. Section 5.1 describes the image data set from COREL in detail. Section 5.2 compares the annotation performance of three MIL-based approaches, Sequential Point-Wise Diverse Density algorithm (SPWDD) [17], ASVM-MIL, and mi-SVM. Finally, we present the results of ASVM-MIL on the benchmark MUSK data sets in Section 5.3.

5.1. Data Description

The data set used for image annotation in this paper is same as the data set used in the experiment of [8]. There are $5,000$ images from 50 Corel Photo CDs in this data set. Each image comes with $4-5$ keywords annotated by Corel employees. $4,500$ images are used for training and the remaining 500 are used for testing. Images are segmented using Normalized-cuts [14]. Only regions larger than a threshold are selected; each image is typically represented by $5-10$ regions sorted by region size. A 33 dimensional low-level feature vector is extracted from each region, which includes region color and standard deviation, region average orientation energy (12 filters), region size, location, convexity, first moment, and ratio of region area to boundary length squared. The vocabulary contains 371 different keywords.

5.2. ASVM-MIL for Image Annotation

We have previously proposed to learn a explicit correspondence between image regions and keywords through a MIL algorithm: SPWDD [17]. After a representative image region has been learnt for a given keyword, the classification problem is addressed using a Bayesian approach. The experiment results show that SPWDD outperforms the Machine Translation model [8] in term of recall and precision on the same annotation data set. This demonstrates that MIL provides an effective and efficient solution for image annotation problems. Readers are referred to [17] for details of SPWDD and its comparison with the Machine Translation model on image annotation.

In this section, we compare three MIL algorithms on image annotation problems: SPWDD, ASVM-MIL, and mi-SVM. In particular, we compare ASVM-MIL with mi-SVM because both are SVM-based MIL algorithms.

ASVM is implemented by modifying the source code of lib-svm [5] and the Gaussian kernel, $K(x, y) = exp(-\gamma\|x - y\|^2)$, is used. We noticed that the training data is extremely unbalanced because there are much more negative image regions than positive image regions for a given keyword. In addition, the number of regions in the data set is more than $40,000$ in total, which requires a extended training time of ASVMs. To resolve both issues, we randomly sampled negative bags to construct a balanced training data set.

To determine the parameters γ and θ for ASVMs, we randomly selected 500 images from the training set to form an independent validation set. Five values are uniformly selected in the range of $[2^{-2}, 2^{2}]$ and $[\theta_0, \theta_1]$, respectively, for γ and θ (25 pairs in total). The pair of parameters that achieves best annotation results on the validation set is chosen as the final one for testing. The mi-SVM is also implemented based on lib-svm using the same kernel function. The kernel parameter γ is determined similarly based on the same validation set as ASVM.

One of the major findings in [17] is that seldom used keywords can not be effectively learnt by MIL methods due to insufficient training examples. Consequently, we select 70 mostly used keywords in the data set and perform our second experiment. The average precision and recall for SPWDD, ASVM-MIL, and mi-SVM on 70 and 30 mostly used keywords are reported in Table 1. The best annotation performance (both recall and precision) is obtained by ASVM. In addition, a closer analysis of precision and recall for the 30 mostly used keywords is provided in Figure 2. The results clearly show that ASVM performs much better than SPWDD and mi-SVM on keywords with diverse visual characters, such as “trees”, “people”, and “rocks”. On other keywords, the three MIL methods have mixed performance.

Finally, Table 2 shows the comparison of the ground truth of three sample images with their annotation results provided by SPWDD, ASVM-MIL, and mi-SVM. Since keywords with similar semantic meanings often have same representative regions, these words usually are included or excluded simultaneously in the final annotation. For example, keywords “cat” and “tiger” have the similar semantic meaning, and both of them appear in the annotation for the first image of Table 2. The third image in Table 2 shows that concepts with similar visual features can hardly be differentiated by region-based annotation systems. Keywords “plants”, “leaf”, and “garden” are wrongly included in the annotation of a “field” image by all three MIL algorithms because they are represented by similar color, texture, and shape features.

5.3. MUSK Data Sets

The MUSK data sets, MUSK1 and MUSK2, are benchmark data sets for MIL. Both data sets consist of descriptions of molecules. Specifically, a bag represents a molecule; instances in a bag represent low-energy confirmations of the molecule. Each conformation is represented by a 166-dimensional feature vector derived from surface properties. MUSK1 contains approximately 6 conformations per molecule on average, while MUSK2 has on average more than 60 conformations in each bag.

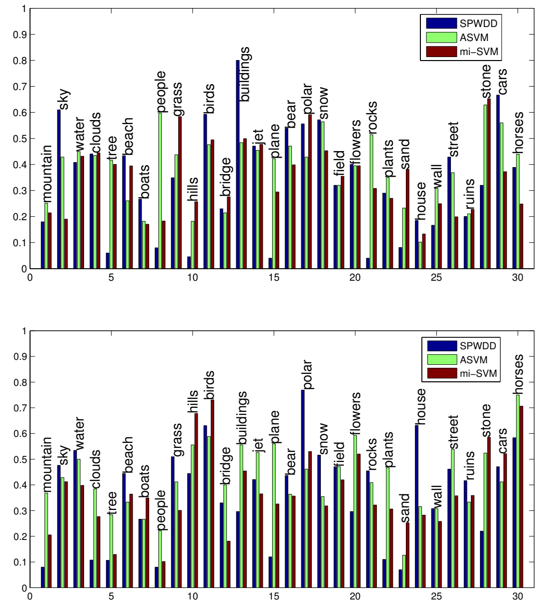


Figure 2. Annotation precision (upper panel) and recall (lower panel) of 30 mostly used keywords by SPWDD, ASVM-MIL, and mi-SVM, respectively

	MUSK 1	MUSK2
ASVM-MIL	**89.1**%	**86.3**%
IAPR	92.4%	89.2%
DD	88.9%	82.5%
mi-SVM	77.9%	84.9%
MI-NN	88.0%	82.5%
DD-SVM	85.8%	91.3%

Table 3. Comparison of 10-fold cross-validation accuracies on MUSK data sets

Again, we use the Gaussian kernel, $K(x, y) = exp(-\gamma\|x - y\|^2)$, for ASVM. Since there is no separate testing set with MUSK, we split the data into ten equal-size groups, eight groups for training, one for validation, and one group for testing. Based on the validation set, the best pair of parameters is grid-searched in the range of $[2^{-2}, 2^{2}]$ and $[\theta_0, \theta_1]$, respectively, for γ and θ. The 10-fold cross-validation accuracy of ASVM on MUSK data sets is reported in Table 3, which also summarizes the performance of five other MIL algorithms in the literature: Interactive APR (IAPR) [7], DD [12], mi-SVM [1], MI-NN [15], and DD-SVM [6].

Table 3 shows that ASVM-MIL achieves very competitive classification accuracy for both MUSK1 and MUSK2 data sets. Note that IAPR has been specifically designed

	SPWDD	ASVM-MIL	mi-SVM
Precision for 70 words	27.31%	**31.19%**	28.36%
Recall for 70 words	35.66%	**39.73%**	35.44%
Precision for 30 words	33.86%	**38.69%**	35.14%
Recall for 30 words	36.88%	**42.70%**	37.89%

Table 1. Average annotation precision and recall for 70 and 30 mostly used keywords by SPWDD, ASVM-MIL and mi-SVM.

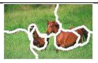
Example	Ground Truth	SPWDD	ASVM-MIL	mi-SVM
	cat tiger water grass	sky cat tiger sunshine field	cat tiger tree polar rocks	cat tiger forest birds rocks
	meadow plants leaf	plants tower field house garden	birds reflection leaf plants nest	house temple grass field plants
	field foals horses mare	plants leaf mountain field tree	tree horses mountain garden window	village horses fox plants field

Table 2. Comparison of the ground truth of three sample images with their annotation results provided by SPWDD, ASVM-MIL, and mi-SVM.

and optimized for the MUSK data sets, the superiority of APR should not be interpreted as a failure. In particular, the result shows that ASVM-MIL greatly outperforms mi-SVM on both data sets even though they are both SVM-based MIL algorithms.

6. Conclusion

In this paper, a novel MIL algorithm: ASVM-MIL is proposed and evaluated for image annotation problems. Our experiments show that ASVM-MIL can greatly improve the image annotation performance, especially for those keywords with diverse appearance. Our experiment results also show that ASVM-MIL runs very competitively with leading MIL methods on the benchmark MUSK data sets.

References

[1] S. Andrews, I. Tsochantaridis, and T. Hofmann. Support vector machines for multiple-instance learning. In *Proc. of the 15th Conference of NIPS*, 2002.

[2] K. Barnad, P. Duygulu, N. Fretias, D. Forsyth, D. Blei, and M. I. Jordan. Matching words and pictures. *Journal of Machine Learning Research*, 3:1107–1135, 2003.

[3] D. Blei and M. Jordan. Modeling annotated data. In *Porc. of the 26th annual intetational ACM SIGIR conference*, 2003.

[4] G. Carneiro and N. Vasconcelos. Formulating semantics image annotation as a supervised learning problem. In *Proc. IEEE CVPR*, 2005.

[5] C. C. Chang and C. J. Lin. Libsvm: a library for support vector machines. 2001.

[6] Y. Chen and J. Wang. Image categorization by learning and reasoning with regions. *Journal of Machine Learning Research*, 5:913 –939, 2004.

[7] T. Dietterich, R. Lathrop, and T. Lozano-Perez. Solving the multipleinstance problem with the axis-parallel rectangles. *Artificial Intelligence*, pages 31 – 71, 1997.

[8] P. Duygulu, K. Barnard, J. de Freitas, and D. Forsyth. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *Proc. of ECCV*, volume 4, pages 97–112, 2002.

[9] J. Fan, Y. Gao, and H. Luo. Multi-level annotation of natural scenes using dominant image components and semantic concepts. In *Proc. of ACM Multimedia*, pages 540–547, 2004.

[10] S. L. Feng, R. Manmatha, and V. Lavrenko. Multiple bernoulli relevance models for image and video annotation. In *Proc. of IEEE CVPR*, volume 1, pages 1002–1009, 2004.

[11] J. Jeon, V. Lavrenko, and R. Manmatha. Automatic image annotation and retrieval using cross-media relevance models. In *Proc. of the 26th Annual Int. ACM SIGIR Conference*, Toronto, Canada, 2003.

[12] O. Maron and T. Lozano-Perez. A framework for multiple-instance learning. In *Proc. of 11th Conference of NIPS*, pages 570–576, 1998.

[13] Y. Mori, H. Takahashi, and R. Oka. Image-to-word transformation based on dividing and vector quantizing images with words. In *First International Workshop on Multimedia Intellegent Storage and Retrieval Management*, 1999.

[14] J. Shi and J. Malik. Normalized cuts and image segmentation. In *Proc. of IEEE CVPR*, Puerto Rico, 1997.

[15] J. Wang and J.-D. Zucker. Solving the multiple-instance problem: a lazy learning approach. In *Proc. of the 17th ICML*, pages 1119 – 1125, 2000.

[16] C. Yang, M. Dong, and F. Fotouhi. Image content annotation using bayesian framework and complement components analysis. In *Proc. of IEEE ICIP*, Genova, Italy, 2005.

[17] C. Yang, M. Dong, and F. Fotouhi. Region-based image annotation through multiple-instance learning. In *Proc. of ACM Multimedia*, Singapore, 2005.