A Novel Multi-instance Learning Algorithm with Application to Image Classification

Xiaocong Xi, Xinshun Xu* and Xiaolin Wang
* Corresponding Author
School of Computer Science and Technology, Shandong University, Jinan 250101, China
xiaocongxi@gmail.com; xuxinshun@sdu.edu.cn; xlwang@sdu.edu.cn

***Abstract*—Image classification is an important research topic due to its potential impact on both image processing and understanding. However, due to the inherent ambiguity of image-keyword mapping, this task becomes a challenge. From the perspective of machine learning, image classification task fits the multi-instance learning (MIL) framework very well owing to the fact that a specific keyword is often relevant to an object in an image rather than the entire image. In this paper, we propose a novel MIL algorithm to address image classification task. First, a new instance prototype extraction method is proposed to construct projection space for each keyword. Then, each training sample is mapped to this potential projection space as a point, which converts the MIL problem into standard supervised learning problem. Finally, an SVM is trained for each keyword. The experimental results on a benchmark data set Corel5k demonstrate that the new instance prototype extraction method can result in more reliable instance prototypes and faster running time, and the proposed MIL approach outperforms some state-of-the-art MIL algorithms.**

I. Introduction

With the rapid increase of digital images on website in recent years, automatic image classification plays a more and more important role in image processing and semantic-based image understanding task. For this task, many approaches have been proposed, which can be roughly divided into two categories, learning-based classification and non-parametric classification. Learning-based classification aims to learn a relevant model between images and keywords which requires an intensive training phase of the classifier, such as Bayesian-based approach [1], SVM-based approach [2, 3, 4], and ensemble learning approaches et al. [5, 6, 7]. This kind of approaches can build a good model between images and keywords, whereas it is an extremely time-consuming to train the model. In contrast with this, non-parametric classification requires no training stage and predicts its classification decision directly on the data. Those methods often rely on Nearest-Neighbor (NN) distance estimation. One of the most popular methods is the kNN approach [8], which classifies an image by the classes of its k nearest images in the data set. Although this category of methods is intuitive and overcomes over-fitting problem of parameters, inferior performance is inevitable due to the poor similarity calculation [4].

In image classification task, we can observe that the specific class is often relevant to an object in an image rather than the entire image. Thus, if we treat the image as a bag and segmented regions in the image as instances, an image with the desired object is labeled as positive, otherwise negative, the image classification task is an intuitive multi-instance learning problem.

Multi-instances learning was first proposed by Dietterich et al. in the context of drug activity prediction [9]. In previous machine learning frameworks, each training sample is treated as a single entity and has a definite label; however, in MIL framework, each training sample is regarded as a bag containing multiple instances. There is no label on the individual instance, only the label of the bag is known. In the binary case, a bag is labeled positive if at least one instance in the bag is positive, and the bag is labeled negative if all the instances in it are negative, the goal of MIL is to classify unseen bags based on the labeled bags. From the above description, we can observe that positive instance (we call it instance prototype) is the key in MIL task. During the past decade many algorithms have been proposed based on it. The early work is axis-parallel concepts [9], the basic idea is to find an axis-parallel rectangle (APR) in the feature space to present target concept, which should contain instance prototypes from positive bags and meanwhile exclude all the instances from negative bags. In this work, three algorithms were proposed to find a hyper-rectangle and evaluated on two real and one artificial data sets for drug activity prediction problem, which showed good performance. In 1998, Maron and Lozano-Perez proposed a general framework called Diverse Density (DD) [10]. This approach is to find the optimal instance prototype in the feature space that is close to at least one instance from every positive bag and meanwhile far away from instances in all negative bags. The optimal instance prototype is defined as the one with the maximum diversity density, which is a measure of how many different positive bags have instances near it, and how far the negative instances are away from it. Later, in 2001, Zhang and Goldman proposed the EM-DD algorithm [11], which combined Expectation Maximization (EM) approach with the DD algorithm to solve this problem. They assumed that the knowledge of which instance determined the label of the bag was a set of hidden variables, and used the EM to find optimal instance prototype. Their method also greatly reduced the complexity of the optimization function and the computational time. With the good performance of SVM, in 2002, Andrews et al. proposed two modified SVM learning methods, mi-

This research was partially supported by the NSFC (61173068) and Shandong Provincial Natural Science Foundation (ZR2009GM021).

SVM for instance-level classification and MI-SVM for bag-level classification [12]. The former is suitable for tasks where users care about instance labels, while the latter is suitable for tasks where only the bag labels are concerned. mi-SVM treats instance labels as unobserved hidden variables subject to constraints defined by their bag labels, and maximizes potential instance prototypes and negative instances margin over the unknown instance labels. In comparison, MI-SVM aims at maximizing the bag margin, which is defined as the margin of the most reliable instance prototype in case of positive bags, or the margin of the least negative instance in case of negative bags. Besides, Chen et al. proposed DD-SVM and MILES methods in [13, 14], their work also learned a collection of instance prototypes, and determined the bag labels based on those instance prototypes.

Due to the important role of instance prototypes in MIL task, in this paper, we propose a new multi-instance learning algorithm. First, a new instance prototype extraction algorithm is proposed to obtain instance prototypes for each keyword. And then, x-means algorithm is adopted to construct projection space, and each training sample is mapped to this potential projection space as a point. This mapping process transforms the MIL task into a traditional supervised learning task. Finally, an SVM is trained for each keyword. We test the proposed approach on the Corel5k data set. The results show that our new instance prototype extraction algorithm can result in reliable instance prototypes and short running time. Moreover, the proposed approach outperforms some state-of-the-art MIL algorithms on image classification task.

The remainder of this paper is organized as follows: Section II gives a specific description of our proposed algorithm. Then, Section III addresses the implementations of our method and presents experimental comparisons with several related methods. Conclusions are given in Section IV.

II. THE PROPOSED ALGORITHM

A. Instance Prototype Extraction

From the perspective of MIL, an image is labeled positive only if this image contains at least one instance prototype for that keyword; otherwise, labeled negative. However, the training data set is very ambiguous in MIL problem. In other words, all the instances in the negative image bags are truly negative instances, but in the positive bags, only a subset of instances are positive, maybe from one instance to all instances in that bag. So how to select instance prototypes from positive bags for a given keyword becomes the basic issue of MIL.

Diverse Density is the first approach to search for the instance prototype. In DD, only one instance with the maximum diversity density is selected, which is a measure of how many different positive bags have instances near that instance, and how far the negative instances are away from that instance. However, in the scenario of image classification, most keywords exhibit a great diversity of visual appearance, e.g., the keyword "sky" often has the blue instance prototype in the beach scene and has the red instance prototype in the sunset scene. So it is improper to use the DD algorithm directly for the image classification task. Several extensions of the DD method have been proposed, aiming at learning more complicated instance prototypes other than a single instance prototype. In [13], Chen and Wang select local maximum of diverse density as instance prototypes; however, it is extremely time consuming. Instead, in [14], Chen et al. select all instances in some bags as instance prototypes which contain much noise. In 2010, Feng and Xu proposed a reinforced DD (reDD) method to search instance prototypes in an efficient and effective way [15], which combined Rahmani and Goldman's method in [16] with Qi and Han's method in [17]. The advantages of such combination lie in the following two folds: First, a more robust DD method is utilized, which is more resistant to the presence of outliers. Secondly, following Rahmani and Goldman's idea, this reinforced DD algorithm can work directly with the MI data, which precludes the need for the multiple starts that are necessary in most existing EM-based algorithms, thus the running speed is improved obviously. However, there are still some weaknesses in this method which will be shown in the following paragraph.

To introduce our proposed method, we use the following notations. Let the labeled bags denote as: $L = \{(B_1, y_1), (B_2, y_2), ...(B_{|L|}, y_{|L|})\}$. Given a specific keyword $w \in V$, according to whether the keyword w is annotated on the image, we further separate the total training set L into positive bags $L^+ = \{B_1^+, B_2^+, ...B_{|L^+|}^+\}$ and negative bags $L^- = \{B_1^-, B_2^-, ...B_{|L^-|}^-\}$, where $B_i^+ = \{x_{ij}^+ | j = 1, 2, ...n_i\}$ is the i^{th} positive bag in L^+, x_{ij}^+ is the j^{th} instance in B_i^+, n_i is the total number of instances for bag i; likewise, there are corresponding same meaning in negative bags. According to Feng and Xu's algorithm [15], for each instance I in positive bags, the DD value of I is defined as:

$$DD(I, L) = \sum_{i=1}^{|L|} Pr((B_i, y_i)|I) \tag{1}$$

Where $Pr((B_i, y_i)|I)$ is a measure of the likelihood that bag B_i receives label y_i given that I belongs to the instance prototypes, the better instance prototype will result the higher $Pr((B_i, y_i)|I)$ value, and thus result higher DD value. However, two observations can be easily made by analyzing (1) as below, which may degrade their performances. To get such observations, we change (1) to its equivalent form:

$$DD(I, L) = \sum_{i=1}^{|L^+|} Pr((B_i, y_i)|I) + \sum_{i=1}^{|L^-|} Pr((B_i, y_i)|I) \tag{2}$$

- Note that in image classification field, the training set (such as Corel5k) is extremely imbalance since there are much more negative bags than positive bags for a given keyword, i.e. $|L^+| \ll |L^-|$. Thus, the second part in (2) will make much more contributions to the value of $DD(I, L)$ than the first part, as a result, the DD value can be totally dominated by the negative bags and the contribution from positive bags is sheathed. However, it is a natural cognition for us that positive bags should

contribute more in finding instance prototypes rather than negative bags.

- Obviously, we also observe that: In the computation of the above DD values, the second part from negative bags accounts for a great proportion in the running time, i.e. $|L^-|/|L|$, this ratio can be extremely high due to the great imbalance in some data set, which makes the algorithm time-consuming.

Based on the aforementioned observations, we modify Feng and Xu's algorithm and propose a new instance prototype extraction algorithm as follows. For each instance I in positive bags, the DD value of I is defined as:

$$DD(I,L) = \sum_{i=1}^{|L^+|} Pr((B_i, y_i)|I) \tag{3}$$

$$Pr((B_i, y_i)|I) = \max_j \{1 - |y_i - \exp(dist^2(B_{ij}, I))|\} \tag{4}$$

$$dist^2(B_{ij}, I) = \sqrt{\sum_{k=1}^{d} (B_{ij}^k - I^k)^2} \tag{5}$$

So, apart from inheriting the advantages in the Feng and Xu's algorithm, our method can achieve the following benefits: (1) A more accurate DD computation method is utilized to avoid the problem that DD value is totally dominated by negative bags due to the great imbalance. (2) The running speed is further improved due to precluding the large amount of negative data.

Once the DD values of all the instances in positive bags are obtained, the next important step is to determine a threshold dd^*_{thres} for the purpose of instance prototypes selection. Here we use a straightforward way to determine the threshold: For each bag in L^+, at least one instance in that bag results in a maximal dd^*, and then find the minimal dd^* as the threshold. Therefore, all the instances satisfying $dd \geq dd^*_{thres}$ will be selected as instance prototypes.

B. Feature Mapping for All Bags

After getting the collection of instance prototypes for a keyword w, we try to explore those instance prototypes and construct projection space. A natural way is using the k-means clustering method. However, due to the granularity effect, it is hard to construct a good performance model. Aiming to solve this problem, x-means method in [18] is employed here to learn instance prototypes. This method can efficiently search the space of cluster locations and number of clusters by optimizing the bayesian information criterion. Many experiments on x-means method have showed that this technique revealed the true number of classes in the underlying distribution and running time was very short. So we apply this method to group those instance prototypes into the best k clusters, then all the central points of k clusters will construct the projection space.

Let $V = \{v_1, v_2, ...v_k\}$ be the projection space of keyword w, where v_i is the i^{th} dimension of projection feature. Using the mapping function in [19], we define $\Phi(B_i)$ as the project feature for bag B_i.

$$\Phi(B_i) = [s(v_1, B_i), s(v_2, B_i), ...s(v_k, B_i)] \tag{6}$$

$$s(v_k, B_i) = \max_j \exp(-\|x_{ij} - v_k\|^2) \tag{7}$$

$$(j = 1, 2, ...n_i)$$

In this way, each bag is mapped into a k-dimensional feature as a point of the projection space. If a bag is positive, the corresponding projection is labeled $+1$; otherwise, labeled -1. Then MIL problem is converted into the traditional standard learning problem. This process can also be interpreted as feature dimension reduction.

C. Image Classification

Using the above method, both training and testing bags are mapped into the projection space as points, so we can use the traditional supervised machine learning to learn from those mapping data for prediction, SVM is adopted in our work to model relations between images and keywords .

We now assume we have $|L|$ labeled mapping data $\{\Phi(B_i), y_i\}$, where $i = 1, 2, ...|L|, y_i \in \{+1, -1\}$. Our intention is to construct an SVM for each keyword with better performance on those mapping data. That is, we need to resolve the following optimization problem:

$$minimize \quad \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{|L|} \xi_i \tag{8}$$

$$s.t. \quad y_i(w^T\Phi(B_i) + b) \geq 1 - \xi_i \tag{9}$$

$$\xi_i \geq 0 \tag{10}$$

In order to further overcome the imbalance problem in the mapping space, we use the same setup in [20], the weight of the positive data is set to $|L|/|L^+|$ and the weight of the negative data is set to $|L|/|L^-|$, let w^* be the optimal solution, then the classifier is defined as:

$$Label(B) = sign(w^{*T}\Phi(B)) \tag{11}$$

The detailed steps of the proposed MIL learning algorithm are summarized in Table I.

TABLE I
THE PROPOSED MIL ALGORITHM

Training stage:
Input: A set of labeled bags L, keyword W
Output: Projection space R and SVM classifier w^*
(a) Partition the training set L into positive bags and negative bags according to the presence of W.
(b) For each instance in positive bags, compute the DD value using (3)-(5), and then select instances with the DD values larger than a threshold as instance prototypes.
(c) Apply x-means algorithm to construct the projection space R, and map all samples into this projection space as points using (6)-(7).
(d) Train a SVM classifier w^* on mapping data.
Predicting stage:
For each test bag, extract its project feature in space R use (6)-(7), and use SVM classifier w^* to predict its label according to (11).

Fig. 1. Some examples in comparison of two instance prototype extraction methods with 4 different keywords (Sky, Tiger, Horses, Formula). The left column is our new proposed instance prototype extraction method, and the right column is Feng and Xu's algorithm.

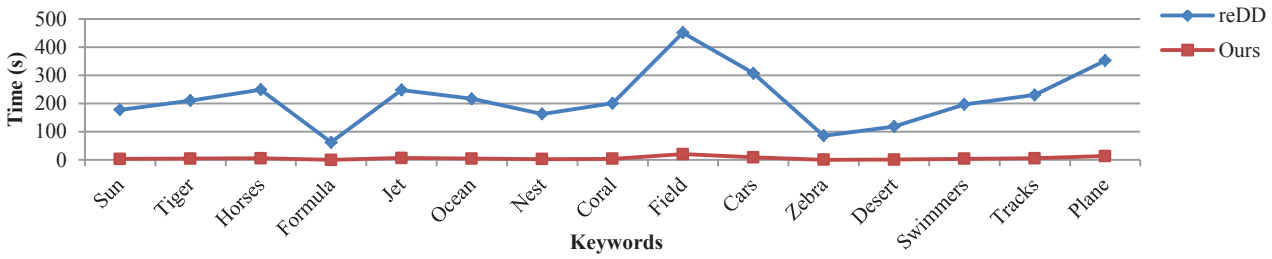

Fig. 2. The computation time with those two instance prototype extraction methods.

III. Experimental Results

A. Data Set

We use the data set Corel5k provided by Duygulu et al. to evaluate our method and compare it with other popular methods [21]. There are 5000 images from 50 CDs, 4500 images are used for training and the remaining 500 are used for testing. Each image is associated with 4-5 keywords to describe the semantic of the image, and is represented by 5-10 regions sorted by region size, regions is produced using normalized-cuts algorithm. A 36 dimensional low-level feature vector is extracted from each region, which includes region color and standard deviation, region average orientation energy, region size, location, convexity, first moment, and ratio of region area to boundary length squared. There are total 374 keywords. To show the effectiveness of the proposed method, we select 15 categories in Corel5k which have great imbalance problem: "Sun," "Tiger," "Horses," "Formula," "Jet," "Ocean," "Nest," "Coral," "Field," "Cars," "Zebra," "Desert," "Swimmers," "Tracks," "Plane."

B. Evaluation

We adopt "one-vs-all" strategy for image classification task, i.e., an SVM is trained to separate a category from all the other categories. We also use the most widely accepted measurement Area Under roc Curve (AUC) as our evaluation. As the name suggests, AUC value is the area size of that part under the roc curve, and describes the probability that a randomly chosen positive sample will be ranked higher than a randomly chosen negative sample. Typically, the AUC values range from 0.5 to 1.0, the larger the AUC value, the better performance.

C. Improvement of Our Instance Prototype Extraction

First of all, to display our instance prototype extraction method visually, we conduct our instance prototype extraction algorithm on the Corel5k data set and compare it with the Feng and Xu's method. Fig. 1 illustrates some detailed examples of the two methods. From the experimental results, we can find that our DD algorithm achieves better results than Feng and Xu's method. For example, instance prototypes extracted by our method are more consistent, and those generated by Feng and Xu's method contain a lot of noise. The computation time of those two instance prototype extraction methods is

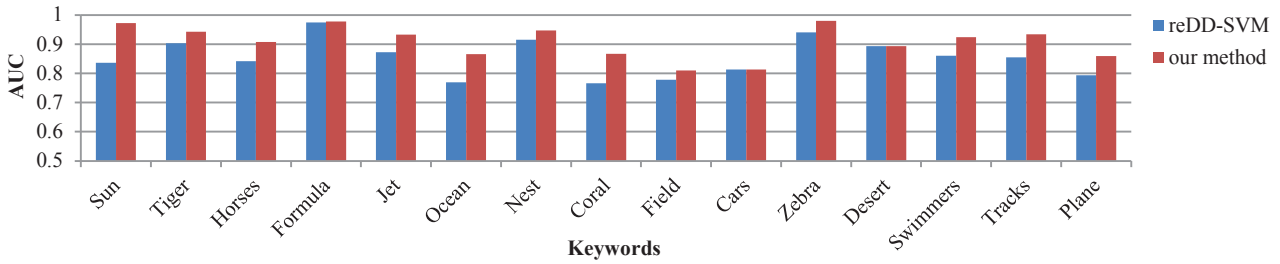

Fig. 3. The detailed AUC value for individual keyword in comparison with our method and reDD-SVM.

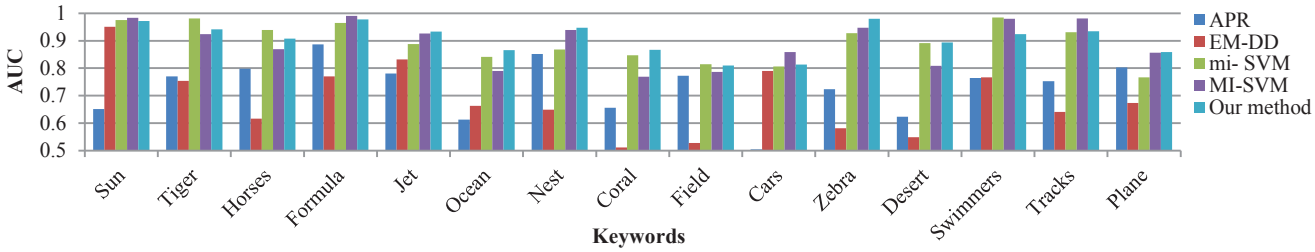

Fig. 4. The detailed AUC value for individual keyword in comparison with our method and four MIL algorithms.

TABLE II
THE AVERAGE AUC IN COMPARISON WITH REDD-SVM

Approach	Avg.AUC
reDD-SVM	0.8540
Our method	0.9084

TABLE III
THE AVERAGE AUC FOR 15 CATEGORIES BY DIFFERENT APPROACHES

Approach	Avg.AUC
APR	0.7303
EM-DD	0.6851
mi-SVM	0.8952
MI-SVM	0.8940
Our method	0.9084

given in Fig. 2. Due to precluding a large amount of negative data in our proposed instance prototype extraction method, the running speed is improved obviously. In summary, our instance prototype extraction method achieves both efficiency and accuracy in finding instance prototypes.

Secondly, we conduct our method to predict the 500 testing images, and replace our instance prototype extraction method with Feng and Xu's algorithm (we call it reDD-SVM) as comparison. All the algorithmic parameters in reDD-SVM are the same as our method. Table II gives the comparison of our method with reDD-SVM in terms of average AUC. Fig. 3 illustrates the detailed results for individual labels. From those experimental results, we can conclude that our method achieves better overall performance with around 6.4% improvement compared with reDD-SVM, which proves that our instance prototype extraction method is more effective in the image classification task.

D. Compare with Various Method

To further show the effectiveness of the proposed method, we conduct experiments to compare with four representative methods for MIL problem: Iterated-discrim APR, EM-DD, mi-SVM, and MI-SVM. Table III gives the comparison results of different methods in terms of average AUC over the 15 labels, while Fig. 4 illustrates the detailed results for each individual labels.

From the experimental results, the following observations can be obtained that our method achieves the best overall performance and obtains around 24.4%, 32.6%, 1.5% and 1.6% improvement compared to iterated-discrim APR, EM-DD, mi-SVM, MI-SVM, respectively, which confirms the effectiveness of the proposed method.

In conclusion, our new instance prototype extraction method achieves both efficiency and accuracy in finding instance prototypes, and our method makes a good contribution to the MIL algorithm for the image classification task.

E. Experiments on Balanced Data

Note that all the above experiments are conducted on imbalanced data of 15 categories in Corel5k, in this part, we will explore whether our proposed method is still effective on balanced data. In order to get such balanced data, here we randomly sample the negative bags as equal to the positive bags on 15 categories. We use this data to train SVM classifiers and classify the 500 testing images. MI-SVM, mi-SVM, reDD-SVM are for comparisons. The final results are shown in Table IV. The experimental results show that: First, compare with former performances in part D, our method, mi-SVM, MI-SVM have a little declines in average AUC respectively, it may

TABLE IV
THE AVERAGE AUC FOR 15 CATEGORIES ON BALANCED DATA

Approach	Avg.AUC
mi-SVM	0.8707
MI-SVM	0.8711
reDD-SVM	0.8838
Our method	0.8791

be caused by the reduced training data generated by balanced data construction. Second, only reDD-SVM gets improved and outperforms the other three methods, so we can conclude Feng and Xu's method is more suitable for balanced data. Third, comparing our method with reDD-SVM, we can find that: For image classification task, adding some negative samples properly in instance prototype extraction part is advisable. However, in our proposed method, we remove all the negative samples in finding instance prototypes, thus all contributions of negative samples are omitted. Therefore, how to use negative samples appropriately and efficiently in the instance prototype extraction part is our future research work.

IV. CONCLUSIONS

In this paper, we propose a novel MIL algorithm based on a new instance prototype extraction method and apply it to the image classification task. With our new proposed instance prototype extraction method, MIL task is transformed into standard supervised machine learning problem efficiently and accurately. The experimental results on a benchmark data set Corel5k demonstrate that our new instance prototype extraction method can obtain more reliable instance prototypes and shorter running time. The results on image classification task also show that the proposed method outperforms some state-of-the-art MIL approaches.

REFERENCES

[1] F.-F. Li, R. Fergus, and P. Perona, "Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories," in *Computer Vision and Image Understanding,* vol. 106, no. 1, pp. 59–70, 2004.

[2] A. Bosch, A. Zisserman, and X. Munoz, "Representing shape with a spatial pyramid kernel," in *Proceedings of ACM International Conference on Image and Video Retrieval,* pp. 401–408, 2007.

[3] M. Marszalek, C. Schmid, H. Harzallah, and J. van deWeijer, "Learning object representations for visual object class recognition," in *Proceedings of IEEE International Conference on Computer Vision, Visual Recognition Challenge workshop,* 2007.

[4] M. Varma and D. Ray, "Learning the discriminative power-invariance trade-off," in *Proceedings of IEEE International Conference on Computer Vision,* pp. 1–8, 2007.

[5] A. Opelt, M. Fussenegger, A. Pinz, and P. Auer, "Weak hypotheses and boosting for generic object detection and recognition," in *Proceedings of IEEE International Conference on Computer Vision,* pp. 71–84, 2004.

[6] X.-S. Xu, X. Xue, and Z.-H. Zhou, "Ensemble multi-instance multi-label learning approach for video annotation task," in *Proceedings of ACM International Conference on Multimedia,* pp. 1153-1156, 2011.

[7] X.-S. Xu, Y. Jiang, L. Peng, X. Xue, and Z.-H. Zhou, "Ensemble approach based on conditional random field for multi-label image and video annotation," in *Proceedings of ACM International Conference on Multimedia,* pp. 1377-1380, 2011.

[8] O. Boiman, E. Shechtman, and M. Irani, "In defense of nearest-neighbor based image classification," in *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition,* pp. 1–8, 2008.

[9] T. G. Dietterich, R. H. Lathrop, and T. Lozano-Prez, "Solving the multiple instance problem with axis-parallel rectangles," in *Journal of Artificial Intelligence,* vol. 89, no. 1–2, pp. 31–71, 1997.

[10] O. Maron and T. Lozano-Prez, "A framework for multiple-instance learning," in *Proceedings of Advances in Neural Information Processing Systems,* pp. 570–576, 1998.

[11] Q. Zhang and S. A. Goldman, "EM-DD: An improved multiple-instance learning technique," in *Proceedings of Advances in Neural Information Processing Systems,* pp. 1073–1080, 2001.

[12] S. Andrews, I. Tsochantaridis and T. Hofmann, "Support vector machines for multiple-instance learning," in *Proceedings of Advances in Neural Information Processing Systems,* pp. 561–568, 2003.

[13] Y. Chen and J. Z. Wang, "Image categorization by learning and reasoning with regions," in *Journal of Machine Learning Research,* vol. 5, no. 8, pp. 913–939, 2004.

[14] Y. Chen, J. Bi and J. Z. Wang, "MILES: Multiple-instance learning via embedded instance selection," in *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 28, no. 12, pp. 1931–1947, 2006.

[15] S. Feng and D. Xu, "Transductive multi-instance multi-label learning algorithm with application to automatic image annotation," in *Journal of Expert Systems with Applications,* vol. 37, no. 1, pp. 661–670, 2010.

[16] R. Rahmani and S. A. Goldman, "MISSL: multiple-instance semi-supervised learning," in *Proceedings of International Conference on Machine Learning,* pp. 705–712, 2006.

[17] X. Qi and Y. Han, "Incorporating multiple SVMs for automatic image annotation," in *Journal of Pattern Recognition,* vol. 40, no. 4, pp. 728–741, 2007.

[18] D. Pelleg and A. Moore, "X-means: Extending k-means with efficient estimation of the number of clusters," in *Proceedings of the International Confence on Machine Learning,* pp. 727–734, 2000.

[19] D. Li and J. Peng, "Object-based image retrieval using semi-supervised multi-instance learning," in *Proceedings of International Congress on Image and Signal Processing,* pp. 1–5, 2009.

[20] E. A. Koen, V. D. Sande, and T. Gevers, "Evaluating color descriptors for object and scene recognition," in *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 32, no. 9, pp. 1582–1596, 2010.

[21] P. Duygulu, K. Barnard, J. D. Freitas, and D. Forsyth, "Object recognition as machine translation: Learning a lexicon for a field image vocabulary," in *Proceedings of European Conference on Computer Vision,* pp. 97–112, 2002.