Journal of Information Hiding and Multimedia Signal Processing ©2019 ISSN 2073-4212
Ubiquitous International Volume **10**, Number **2**, March **2019**

Research on Multi-instance Learning for Semantic Image Analysis

Dongping Tian

Institute of Computer Software
Baoji University of Arts and Sciences
No.1 Hi-Tech Avenue, Hi-Tech District, Baoji, Shaanxi 721013, P.R. China

Institute of Computational Information Science
Baoji University of Arts and Sciences
No.44 Baoguang Road, Weibin District, Baoji, Shaanxi 721016, P.R. China
tiandp@ics.ict.ac.cn, tdp211@163.com

Received August 2018; revised December 2018

Abstract. *Semantic image analysis is an active topic of research in computer vision and pattern recognition. In the last two decades, a large number of works on semantic image analysis have emerged, among which the multi-instance learning (MIL) is one of the most commonly used methods due to its theoretical interest and its applicability to real-world problems. However, compared with various MIL methods and their corresponding applications in the field of semantic image analysis, there is a lack of surveys or review researches about MIL related studies. So the current paper, to begin with, elaborates the basic principles of multi-instance learning, subsequently summarizes it with applications to semantic based- image annotation, image retrieval and image classification as well as several other related applications comprehensively. At length, this paper concludes with a summary of some important conclusions and several potential research directions of MIL in the area of semantic image analysis for the future.*
Keywords: MIL, Image annotation, CBIR, Image classification, PLSA, GMM, SVM

1. **Introduction.** With the explosive growth of the world wide web and rapidly growing number of available digital color images, much research effort is devoted to the development of efficient semantic image analysis systems. The past few years have witnessed a considerable progress in this area, yet, as a field, automatic image annotation (AIA) is still in its infancy, facing many challenges and limitations. One of the main handicaps is the well known semantic gap between low-level visual features and high-level semantic concepts. Fortunately, a huge number of advanced machine learning techniques have been proposed in the literature as a promising solution to fill the semantic gap. As a pioneer work, Duygulu et al.[1] proposed the translation model (TM) to treat AIA as a process of translation from a set of blob tokens, obtained by clustering image regions, to a set of keywords. Jeon et al.[2] presented the cross-media relevance model (CMRM) to annotate image, assuming that the blobs and words are mutually independent given a specific image. Subsequently, CMRM was improved through the continuous space relevance model [3], multiple Bernoulli relevance model (MBRM)[4] and dual cross-media relevance model [5]. As latent aspect models, Monay et al. put forward a series of probabilistic latent semantic analysis (PLSA) models for AIA [6-8], among which PLSA-MIXED [6] learned

a standard PLSA model on a concatenated representation of the textual and visual features, while PLSA-WORDS or PLSA-FEATURES [7,8] allowed modeling of an image as a mixture of latent aspects that was defined either by its textual captions or by its visual features for which the conditional distributions over aspects were estimated from one of the two modalities only. In our previous work [9,10], the unified two-stage refining image annotation methods have been proposed by integrating PLSA with random walk and max-bisection models respectively. In addition, Xu et al.[11] came up with a regularized latent Dirichlet allocation model for tag refinement, which facilitated the topic modeling by exploiting both the statistics of tags and visual affinities of images in the corpus. Meanwhile, several nearest-neighbor based methods have been proposed in recent years [12,13]. Alternatively, it should be noted that the multi-instance learning, as another kind of supervised learning method, has also been widely used in the community of computer vision [14-62,66-77].

As briefly reviewed above, most of these approaches can achieve promising performance and motivate us to explore better semantic image analysis methods with the help of their excellent experiences and knowledge. So in this paper, we provide a survey of MIL that related to the semantic image analysis in the last decade rather than a concrete image annotation method. The primary purpose of this paper is to illustrate the effectiveness of MIL and how to further improve its applications in the field of computer vision and pattern recognition. The remainder of this paper is organized as follows. Section 2 elaborates the basic principles of MIL method. In section 3, the MIL with applications to image annotation, image retrieval, image classification and some other applications are reviewed comprehensively. Finally, we conclude this paper in section 4 with a summary of some important conclusions and highlight the potential research directions of MIL in semantic image analysis for the future.

2. **Multi-instance Learning.** Multi-instance learning (MIL)[1][14] is a variation of supervised learning, where the task is to learn a concept given positive and negative bags of instances. Each bag may contain many instances, but a bag is labeled positive even if only one of the instances in it falls within the concept. On the contrary, a bag is labeled negative only if all the instances in it are negative. Hence, learning focuses on finding the actual positive instances in the positive bags. In other words, the goal of MIL is to generate a classifier that can classify unseen bags correctly. Formally, the task of MIL is to learn a function as follows [15]:

$f_{MIL} : 2^{\chi} \to \{-1, +1\}$ from a given data set $\{(X_1, y_1), (X_2, y_2), \cdots, (X_m, y_m)\}$, where $X_i \subseteq \chi$ denotes a set of instances $\{x_1^{(i)}, x_2^{(i)}, \cdots, x_{n_i}^{(i)}\}$, $x_j^{(i)} \in \chi(j = 1, 2, \cdots, n_i)$, $y_i \in \{-1, +1\}$ is the label of X_i.

In the context of automatic image annotation, an image is usually described by multiple semantic labels (or keywords) and these labels are often highly related to respective regions rather than the entire image itself. As a result, it is unclear which region in an image is associated with which class label assigned to the image. From this sense, the problem of AIA can be effectively solved by a more rational and natural strategy, i.e., the multi-instance learning method. To summarize, the task of AIA can be formulated as a MIL problem based on the following two aspects. On one hand, each segmented region is treated as an instance and all of them are grouped to form an image as a bag of instances. On the other side, at least one label should be assigned on each bag. Given an image labeled by keyword w_i, it is expected that at least one region will correspond to w_i even if segmentation may not be perfect. Hence, the image annotation problem is in essence

[1]https://prlab.tudelft.nl/david-tax/mil.html

identical to MIL setting. One way to solve MIL problem is to examine the distribution of these instances and look for an instance that is close to all instances in the positive bags whereas far away from those from negative bags. In other words, we should search for a point where there is a high diverse density (DD) of positive instances. As is well known, the point-wise diverse density (PWDD) approach [14] was the first probability model of MIL, which has been widely used to look for the target concept. The DD method measured a co-occurrence of similar instances from different bags with the same label. A feature point with large DD value indicated that it was close to at least one instance from every positive bag and far away from every negative instance. A gradient ascent method was in general adopted to search the instance feature space for points with high diverse density. Particularly, PWDD was very useful for image annotation since it can return the most representative regions for a keyword, which makes it possible to explicitly observe the correspondence between regions and keywords.

3. MIL for Semantic Image Analysis.

From the literature, it can be seen that MIL has become an important topic in the pattern recognition community, and many solutions to this problem have been proposed until now. Thus in this section, MIL for semantic image analysis will be summarized from the aspects of image annotation, image retrieval, image classification and other specific applications respectively. It is not possible to list all the existing MILs. Instead, we mainly focus on various MIL methods associated with semantic image analysis and attempt to look into them through a unified view.

3.1. MIL for image annotation.

Image annotation has been an active topic of research in computer vision for decades due to its potentially large impact on both image understanding and web image search. To be specific, automatic image annotation (AIA) refers to a process to automatically generate textual words to describe the content of a given image, which plays a crucial role in semantic based image retrieval. In this subsection, we will review some pioneer works for automatic image annotation by using MIL related models. Note that during the past decade, many MIL algorithms have been proposed for AIA [16-31]. In the work of [16], a MI-SVM method was developed for automatic image annotation, in which the instance-based (i.e., region-based) image features were iteratively fed into the SVM until no updates for all the positive training images, subsequently the converged instance-based features were used to annotate the unseen images. In the meanwhile, Andrews et al.[17] presented mi-SVM and MI-SVM algorithms for instance-level classification and bag-level classification respectively by modifying the SVM formation. Note that mi-SVM explicitly treats the label instance labels y_i as unobserved hidden variables subject to constraints defined by their bag labels Y_I, whose goal is to maximize the usual instance margin jointly over the unknown instance labels and a linear or kernelized discriminant function, given below:

$$\begin{aligned} &\min_{\{y_i\}} \min_{w,b,\zeta} \frac{1}{2}\|w\|^2 + C\sum_i \zeta_i \\ &s.t. \forall i : y_i\left(\langle w, x_i\rangle + b\right) \geqslant 1 - \zeta_i, \zeta_i \geqslant 0 \\ &\sum\nolimits_{i\in I} \frac{y_i + 1}{2} \geqslant 1, \forall I : Y_I = 1; y_i = -1, \forall I : Y_I = -1 \end{aligned} \tag{1}$$

where the second part of the constraint enforces the relations between instance labels and bag labels. In comparison, MI-SVM aims at maximizing the bag margin, which is defined as the margin of the "most positive" instance in case of positive bags, or the margin of

the "least negative" instance in case of negative bags, given as:

$$\begin{aligned} &\min_{w,b,\zeta} \frac{1}{2}\|w\|^2 + C\sum_I \zeta_I \\ &s.t. \forall I : Y_I \max_{i\in I}(\langle w, x_i\rangle + b) \geqslant 1 - \zeta_I, \zeta_I \geqslant 0 \end{aligned} \tag{2}$$

Note that in mi-SVM the margin of every instance matters, and one has the freedom to set the instance label variables under the constraints of their bag labels to maximize the margin. In comparison, in MI-SVM only one instance per bag matters since it determines the margin of the bag. The former is suitable for tasks where users care about instance labels whereas the latter is suitable for tasks where only the bag labels are concerned. Both methods are implemented using mixed integer quadratic programming. However, unlike the standard SVM, they would lead to the non-convex optimization problems that suffer from local minima. As a result, Gehler et al.[18] made use of deterministic annealing to solve this non-convex optimization problem, and proposed AL-SVM method which could find better local minima of the objective function. On the other hand, note that as the pioneer work of MIL for region-based image annotation, Yang et al.[19] proposed to learn an explicit correspondence between image regions and keywords through the sequential PWDD multi-instance learning. Followed by they modeled automatic image annotation as a problem of image classification with the help of Bayesian framework. In order to find an optimal nonlinear decision boundary for each concept, they developed the asymmetrical support vector machine-based MIL, which extended the conventional SVM in the MIL framework by introducing asymmetrical loss functions for false positives and false negatives [20]. In the work [21], a generalized discriminative MIL was put forward for semantic concept detection by fusing both the expressive power of generative models and the advantage of discriminative training into the MIL setting. Afterwards Zhao et al.[22] introduced the minimum reference set into MIL to construct a novel automatic image annotation scheme, in which the positive instances (i.e. regions in images) embedded in the positive bags (i.e. images) could be picked out via reliable inferring for a concept. Feng et al.[23] formulated image annotation under the MIL framework described in reference [19] and presented an improved Citation-kNN (ICKNN) MIL algorithm for AIA. The main difference between PWDD and ICKNN lies in that the latter algorithm avoids learning the target instance (region) to represent a given keyword from the collection of training bags (images) and the keywords are annotated on the whole image instead of image regions. In other words, the testing bag's labels are directly decided by its neighbor training bags. Followed by they put forward a reinforced DD method to search instance prototypes in an efficient and effective way (abbreviated as TMIML) to solve the issue of automatic image annotation [24], which combined the methods in [25] and [26]. The advantages of such combination mainly lie in twofold: First, a more robust DD method was utilized, which is more resistant to the presence of outliers. Second, following the idea of [25], this reinforced DD algorithm can work directly with the MI data, which precluded the need for the multiple starts that are necessary in most existing EM-based algorithms, thus the running speed was markedly improved. Compared with previously MIML algorithm [15], the TMIML framework is much more effective due to the fact that the large amount of unlabeled samples were taken into account to resolve the small samples problem that often appears in the context of image auto-annotation task.

Besides, Wang et al.[28] presented a decoupled two stage solution to the MIL problem by applying a modified random walk on a graph process to infer the positive instances in each positive bag. Combined with the support vector machine classifier, this algorithm can efficiently decouple the inferring and training stages and convert MIL into a supervised

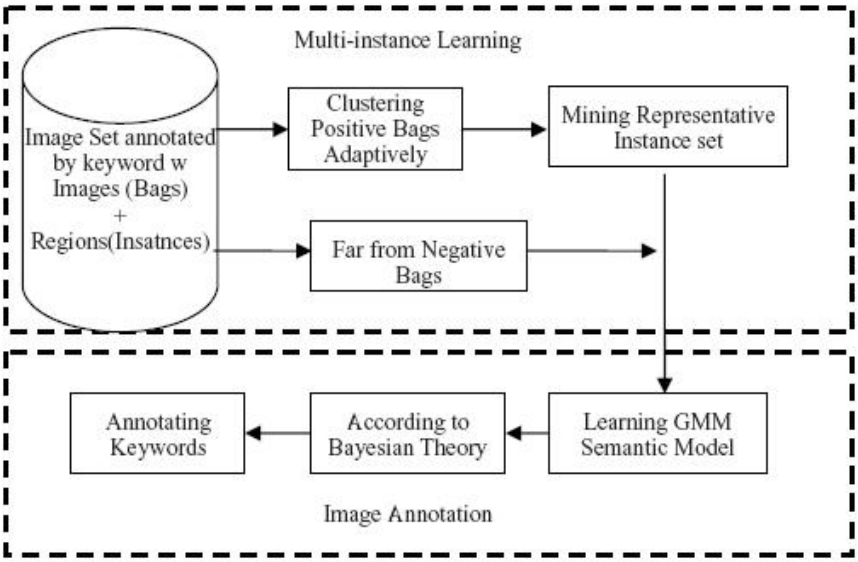

FIGURE 1. Framework of the MIL-based AIA model

learning problem. A recent work by Xue et al.[27] came up with a novel image annotation method based on MIL, in which the input image was segmented and can be viewed as a bag of instances (regions). The global visual features of the entire image and the local features of the regions were extracted to capture coarse and fine patterns respectively. At the same time, Zhu et al.[29] presented a multi-instance learning based AIA model (as illustrated by Fig. 1), in which each keyword was analyzed hierarchically in low-granularity-level under the framework of MIL. In more recent work [30], Nguyen et al. proposed multi-modal multi-instance multi-label latent Dirichlet allocation (M^3LDA) for image annotation, where the model consisted of a visual-label part, a textual-label part and a label-topic part. Specifically, the visual-label and textual-label parts were devoted to the mappings from the visual and textual spaces to the label space while the label-topic part helped to capture the label relationships. The basic idea behind M^3LDA is that the topic decided by the visual information and the topic decided by the textual information should be consistent so as to result in correct label assignment. In [31], a two-stage MIL algorithm was developed for automatic image annotation. To be specific, the affinity propagation (AP) clustering technique was performed on the instances both in positive and negative bags to identify the candidates of the positive instances and to initialize the maximum searching of DD likelihood in the first stage. In the second stage, the most positive instances were selected out in each bag to simplify the computing procedure of DD likelihood.

3.2. **MIL for image retrieval.** In the past years, content-based image retrieval (CBIR) has been one of the most hot research topics in computer vision. Much work has been done in applying MIL to localized content-based image retrieval since the CBIR fits well the MIL framework as an image can be seen as a bag comprised of smaller regions/patches (i.e., instances). Given a query for a particular object, one may be interested in deciding only whether the image contains the queried object or not, instead of solving the more involved problem of labeling every single patch in the image. As the pioneer work of MIL for image retrieval, Maron et al.[32] firstly formulated CBIR as a multiple instance learning problem. In their framework, each image was deemed as a labeled bag with

multiple instances and the segmented regions in the images corresponded to the instances in the bags. Specifically, they developed the DD method to solve the MIL problem by converting the goal of MIL to a maximization problem. That is, with the assumption of n labeled bags and the hypothesis t, the DD value is calculated as:

$$DD(t) = \prod_{i=1}^{n} Pr(B_i, l_i|t) = \prod_{i=1}^{n} (1 - |l_i - Label(B_i|t)|) \tag{3}$$

$$Label(B_i|t) = \max_j \{exp[-\sum_{d=1}^{m} (s_d(B_{ijd} - t_d))^2]\} \tag{4}$$

where B_i denotes the i-th bag, l_i denotes the actual label of the i-th bag, B_{ij} is the j-th instance of bag i, B_{ijd} represents the feature value of instance B_{ij} on dimension d, S_d denotes the value of feature weight vector S on dimension d, t_d denotes the value of t on dimension d, n denotes the number of instances, and m denotes the number of features. The maximization of Eq.(1) is to find the optimum t that leads to the maximum DD value for representing the user's interest in the feature space.

In [33], Yang et al. extended the diverse density algorithm and applied it to content-based image retrieval based on a complex bag generator by MIL algorithm. Followed by Zhou et al.[34] also applied diverse density to CBIR. It is noteworthy that they developed a bag generator (ImaBag) which was derived from a SOM-based image segmentation technique. Experiments showed that the performance of ImaBag is better than that proposed in [33], but worse than that of Maron and Ratan's [32], when they were coupling with diverse density. Zhang et al.[35] combined expectation-maximization (EM) with DD to propose an algorithm named EM-DD to improve the annotation speed and facilitate the scale-up to large data sets. However, the MIL was degraded to a single instance learning since EM only estimated one instance that was responsible for the label of the bag. Subsequently, Zhang and Chen [36] proposed an approach based on one-class support vector machine to solve MIL problem in the region-based CBIR. This is an area where a huge number of image regions are involved. For the sake of efficiency, a genetic algorithm based clustering method was adopted to reduce the search space in conjunction with the relevance feedback technique was incorporated to provide progressive guidance to the learning process. In [37], Yuan et al. formulated region-based image retrieval as a MIL problem and proposed MI-AdaBoost algorithm to solve it. To be specific, this approach first mapped each bag into a new bag feature space using a certain set of instance prototypes and then employed AdaBoost to select the bag features and built classifiers simultaneously.

In spite of many MIL methods applied to CBIR, most of them only have a supervised manner using bag-level labels instead of the information of unlabeled data which do not belong to any labeled bag. In view of this, a semi-supervised SVM framework of MIL algorithm was developed for localized content-based (object-based) image retrieval [38], whose goal is to rank all the images in the database according to the object that users want to retrieve. In [39], a multiple-instance semi-supervised learning (MISSL) was proposed to solve the object-based image retrieval problem. Unlike the loosely coupled manner exhibited in [25], a graph-based multiple-instance learning (GMIL) model was developed based on the regularization framework of MISSL by explicitly taking into account labeled data, semi-labeled data and unlabeled data simultaneously to propagate information on a graph. Here, it should be noted that the object-based image retrieval is related but different from the concept of region-based image retrieval [39]. Besides, a self-taught multiple-instance learning technique was presented to deal with learning from a limited number of ambiguously labeled examples [40], which used a sparse representation for examples belonging to different classes in terms of a shared dictionary derived from the unlabeled data. Particularly, the sparse representation can be optimized under the MIL

setting to both construct high-level features and unite the data distribution. Similar attempts have also been made by Li et al.[41], who utilized relevant and irrelevant training web images rather than image regions to generate bags as well as instances for MIL problem formulation. They constructed a new model called MIL-CPB to effectively exploit the constraints that each positive bag contained at least a portion of positive instances on positive bags and predicted the labels of test instances (images). What's more, they also developed a progressive scheme called progressive MIL-CPB (PMIL-CPB) to further improve the retrieval performance by iteratively partitioning the top-ranked training web images from the current MIL-CPB classifier to construct more confident positive bags and then added these new bags as training data to learn the subsequent MIL-CPB classifiers. In the meanwhile, the LSASVM-MIL model [42] was formulated for image retrieval based on the latent semantic analysis (LSA) and support vector machine (as shown in Fig. 2). Specifically, a LSA based method was first utilized to convert bags in the MIL problem into a single representation vector, and then combining with SVM in the framework of a MIL algorithm for image retrieval. Experiments on Corel datasets validated the effectiveness and efficiency of this model.

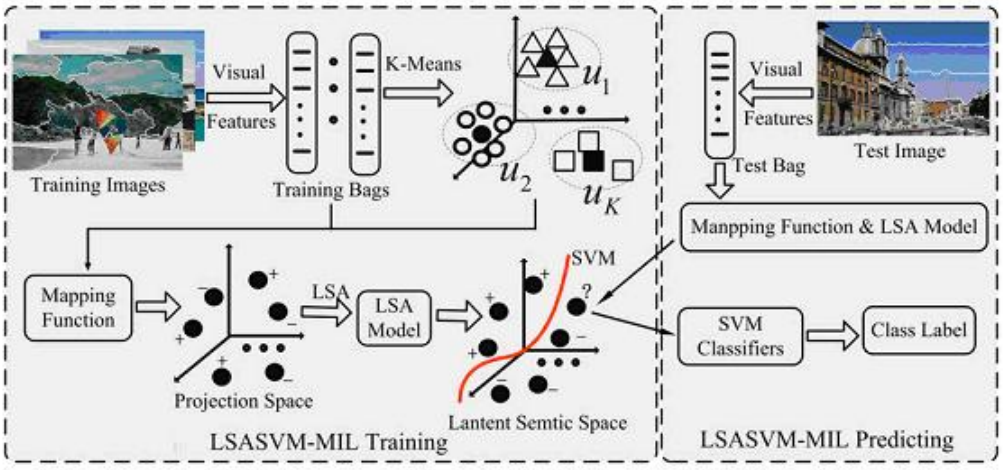

FIGURE 2. Framework of the LSASVM-MIL model

In subsequent work [43], a semi-supervised multi-instance learning (SSMIL) was proposed for localized content-based image retrieval based on PLSA and transductive support vector machine (abbreviated as PLSA-SSMIL), note that in which the latent topic feature was extracted by PLSA rather than other methods in order to better represent the semantic of bag in image retrieval, the semi-supervised TSVM was used to train the classifier so as to take advantage of a large number of unlabeled images to improve the classifier performance, in other words, the small sample learning problem can be well resolved. In addition, the labels were assigned to image rather then region that can greatly improve the efficiency of hand-labeled the training samples. Fig. 3 illustrates the framework of the PLSA-SSMIL model. On the other hand, motivated by the fact that a bag label is solely determined by the instance that has the highest confidence toward the positive class, Kim et al.[44] formulated the bag class likelihood as the sigmoid function over the maximum Gaussian processes latent variables on the instances. By marginalizing out the latent variables, a nonparametric, nonlinear probabilistic model could be obtained that fully respected the bag labeling protocol of the MIL. In more recent work [45], a multiple-instance learning based decision neural network (MI-BDNN) was developed to

attempt to bridge the semantic gap in CBIR. MI-BDNN considered the image retrieval problem as a MIL problem, where a user's preferred image concept is learned by training MI-BDNN with a set of exemplar images, each of which is labeled as conceptual related (positive) or conceptual unrelated (negative) image. Meanwhile, a MIL was put forward based on the representative instances and feature mapping (RIFM-I)[46] for CBIR, experiments showed that RIFM-I could result in superior retrieval performance as well as distinguishing some easily confused categories quite well.

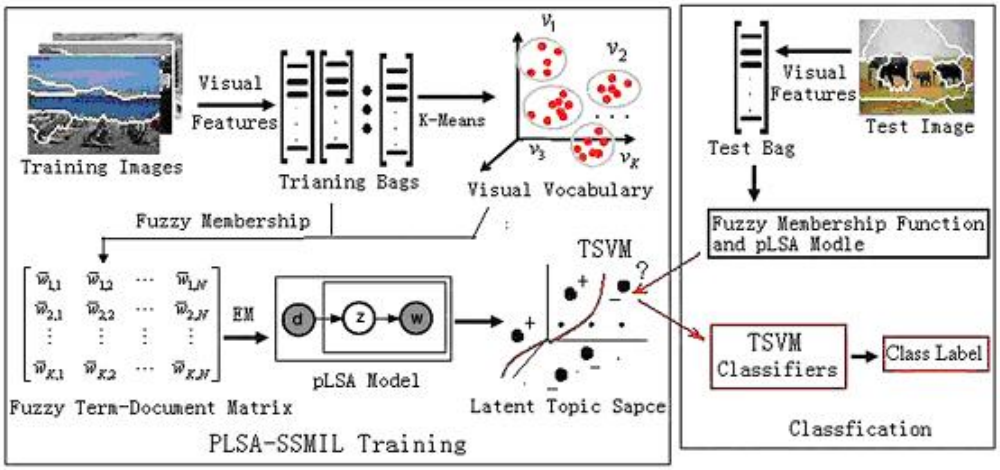

FIGURE 3. Framework of the PLSA-SSMIL model

3.3. **MIL for image classification.** Image classification is an important research topic due to its potential impact on both the image processing and understanding. However, it actually becomes a challenge problem due to the inherent ambiguity of image-keyword mapping. From the viewpoint of machine learning, image classification fits the 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. So far there has been much work on applying MIL to the task of image classification. As a classical work, Chen et al.[47] developed a DD-SVM for mapping every bag to a point in a new feature space defined by the instance prototypes selected from local maxima of DD function, and then a SVM was trained based on the bag features. Note that the maximum margin formulation of MIL in the bag feature space was given as the following quadratic optimization problem:

$$\alpha^* = \underset{\alpha_i}{argmax} \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{l} y_i y_j \alpha_i \alpha_j K(\phi(B_i), \phi(B_j)) \tag{5}$$

$$s.t. \begin{cases} \sum_{i=1}^{l} y_i \alpha_i = 0 \\ 0 \leqslant \alpha_i \leqslant C, i = 1, \cdots, l. \end{cases}$$

Here, it is worth noting that the representation feature of DD-SVM was very sensitive to noise and could easily incur very high computation cost. As a consequence, MILES [48] method exploited all the instances from the training bags instead of the prototypes used with DD-SVM to construct a new feature space. Specifically, MILES mapped each bag into a feature space defined by the instances in the training bags via an instance similarity measure, and then a 1-norm SVM was applied to build the bag level classifiers for image

classification. Another interesting work has been developed by Zhou et al.[15], who first analyzed the relationship between multi-instance and multi-label learning (MIML) and subsequently proposed two corresponding methods, i.e., MIML-Boost and MIMLSVM with application to scene classification. Both of the two algorithms attempted to convert MIML into a traditional supervised learning problem. MIMLSVM and MIML-Boost worked by degenerating a MIML task to a simplified supervised learning task using single-instance multi-label or multi-instance single-label approaches as bridges. Fig. 4 illustrates the differences among these learning frameworks concisely, and for more details please refer to literature [15].

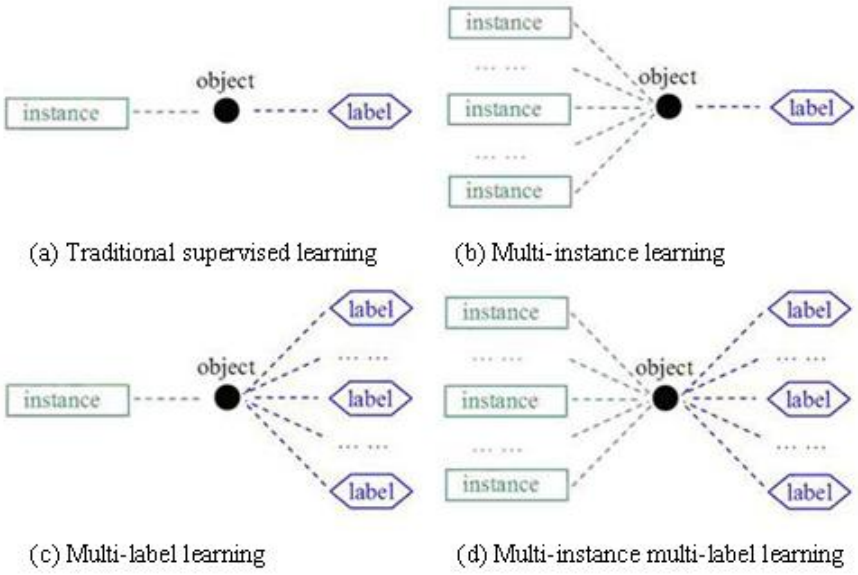

FIGURE 4. Four different learning frameworks

As can be seen from the literatures, the single-instance supervised learning algorithm can be adapted to multi-instance learning as long as its focus is shifted from the discrimination on the instances to the discrimination on the bags. In actual fact, most current MIL algorithms can be viewed as going along this way, which is adapting single-instance learning algorithms to the multi-instance representation. In [49], an EM based learning algorithm was proposed to provide a comprehensive procedure for maximizing the measurement of diverse density on the given multiple instances. In essential, this method converted the multi-instance problem into a single-instance treatment by using EM to maximize the instance responsibility for the corresponding label of each bag. Recently, Li et al.[50] proposed a new image multi-instance bag generating method that modeled an image with a Gaussian mixture model (GMM). Note that the generated GMM was treated as a multi-instance bag, the agglomerative information bottleneck clustering was employed to transform the MIL problem into single-instance learning problem and ensemble learning was involved to further enhance classifiers' generalization ability for image classification. In literature [51], a MIL algorithm was constructed to address image classification involving three steps, i.e., a new instance prototype extraction method was proposed to construct projection space for each keyword, each training sample was mapped to this potential projection space as a point and a SVM was trained for each keyword to implement image classification. It is worth noting that the proposed new

instance prototype extraction algorithm can be formulated as follows. For each instance I in positive bags, the DD value of I is defined as below, where $|L^+|$ denotes the number of positive bags for a given keyword, $Pr((B_i, y_i)|I)$ is a measure of the likelihood that bag B_i receives label y_i given that I belongs to the instance prototypes.

$$DD(I, L) = \sum_{i=1}^{|L^+|} Pr((B_i, y_i)|I) \tag{6}$$

$$\Pr((B_i, y_i)|I) = \max_j \{1 - |y_i - exp(dist^2(B_{ij}, I))|\} \tag{7}$$

$$dist^2(B_{ij}, I) = \sqrt{\sum_{k=1}^{d} \left(B_{ij}^k - I^k\right)^2} \tag{8}$$

In more recent work [52], SCCE-MIL was presented for image categorization by combining sparse coding and classifier ensemble strategy under the MIL framework, which not only possessed the good feature representation ability of sparse coding but also utilized the power of ensemble paradigms to achieve strong generalization ability. In addition, MKSVM-MIL was developed based on the affinity propagation and multiple-kernels support vector machine for image classification [53]. In the meanwhile, MI-NSVM was proposed based on nonparallel classifiers for image classification [54].

3.4. **MIL for other applications.** Multiple-instance learning has been a popular topic in the study of pattern recognition for years due to its usefulness for such tasks as drug activity prediction and image annotation, image retrieval and image classification. Apart from the content of MIL for semantic image analysis described above, many multi-instance learning algorithms have also been intensively studied and applied in many other applications during this decade, such as robot control [55,56], medical image semantic analysis [57-60], video caption detection [61,62], semantic web applications [63-65], object/event detection and tracking [66-71] as well as some specific applications [72-77], etc. Besides, we refer the reader to two recent works [78,79] for more details on a comprehensive and extensive literature survey on multiple-instance learning. In particular, a book on MIL has been recently published [80], which discusses most of the tasks such as classification, regression, ranking and clustering along with the associated methods, as well as the data reduction and imbalanced data. For more details of them please refer to the corresponding literature. In the following, several MIL related semantic image analysis approaches involved in this paper are concisely summarized in Table 1, mainly including the methods adopted and the test datasets employed.

Note:

1. Object image database: *www.sony.com*, *www.flycontinental.com*, *www.delta-air.com*, *www.avis.com*, *www.bicycle.com*, *www.jcpenney.com*, *www.jcrew.com*, *www.ritzcamera.com*, *www.sears.com*, comprising 228 images from 19 different categories.
2. SIVAL database: *www.cs.wustl.edu/sg/multi-inst-data/*.
3. MSRC database: *http://research.microsoft.com/en-us/projects/objectclassrecognition/*.
4. NUS-WIDE: *http://lms.comp.nus.edu.sg/research/NUS-WIDE*.
5. 500-image dataset: cats & dogs, flowers, mountains, planes and buildings, each type contains 100.

4. **Conclusions and Future Work.** Multi-instance learning has become an active area of investigation in computer vision and machine learning since it was first formulized in the context of drug activity prediction. In this work, we present a comprehensive survey on MIL related studies in semantic image analysis, especially from the aspects of image annotation, image retrieval, image classification and several other applications respectively to complement a very small number of existing MIL surveys in literature.

Table 1. Summary of MIL related semantic image analysis models

Sources	**Methods adopted**	**Image dataset applied**
Zhou et al.[15]	MIMLSVM, MIMLBOOST	COREL Dataset
Andrews et al.[17]	MIL, SVM	COREL Dataset
Gehler et al.[18]	MIL, SVM	COREL/MUSK Datasets
Yang et al.[19]	MIL, Bayesian classifiers	COREL Dataset
Yang et al.[20]	Asymmetrical SVM-MIL	COREL/MUSK Datasets
Gao et al.[21]	GDMIL	TRECVID 2005
Zhao et al.[22]	MIL, Minimum reference set	COREL Dataset
Feng et al.[23]	ICKNN MIL	COREL Dataset
Feng et al.[24]	TMIML	COREL Dataset
Xue et al.[27]	MIL, Structural max-margin	COREL/MSRC Datasets
Wang et al.[28]	MIL, SVM, Random walk	COREL/MUSK/TREC9 Datasets
Zhu et al.[29]	MIL, GMM	COREL Dataset
Nguyen et al.[30]	M3LDA	COREL/ImageCLEF Datasets
Xia et al.[31]	MIL, AP clustering	MSRC/NUS-WIDE Datasets
Maron et al.[32]	MIL	COREL Dataset
Yang et al.[33]	MIL	COREL/Object Image Databases
Zhou et al.[34]	MIL, SOM neural network	500-Image Dataset
Zhang et al.[35]	MIL	COREL/SIMPLIcity Datasets
Zhang et al.[36]	MIL, SVM, Relevance feedback	COREL Dataset
Yuan et al.[37]	MI-AdaBoost	COREL/MUSK Datasets
Zhang et al.[38]	SSMIL, UP-SSMIL	SIVAL Dataset
Wang et al.[39]	GMIL, GMIL-M	SIVAL/MUSK Datasets
Qiao et al.[40]	STMIL	COREL/SIVAL Datasets
Li et al.[41]	MIL-CPB, PMIL-CPB	NUS-WIDE/Google Datasets
Li et al.[42]	MIL, SVM, LSA	COREL Dataset
Li et al.[43]	SSMIL, TSVM, PLSA	COREL Dataset
Kim et al.[44]	GPMIL	COREL/MUSK/SIVAL Datasets
Xu [45]	MI-BDNN	COREL/SIVAL Datasets
Chen et al.[48]	MILES	COREL/Caltech/MUSK Datasets
Pao et al.[49]	MIL	OTHER Dataset
Li et al.[50]	MIL, GMM	COREL Dataset
Xi et al.[51]	MIL, SVM	COREL Dataset
Song et al.[52]	SCCE-MIL	COREL Dataset
Li et al.[53]	MKSVM-MIL	COREL Dataset
Amores [78]	MIL, SVM	COREL Dataset

The primary purpose of this paper is to illustrate the pros and cons of MIL combined with a great deal of existing researches as well as to point out the promising research directions of multi-instance learning for semantic image analysis in the future.

A lot of very interesting topics have not been included here but would be worth exploring more in depth in the future. It should be noted that the following several issues remain to be investigated. First, the most serious problem encumbering the advance of MIL is that there is only one popularly used real-world benchmark data, i.e., the Musk data sets. Although some application data have been used in some works, they can hardly act as benchmarks for some reasons. So how to build some publicly available challenging datasets that can estimate the performance of MIL pretty well is a worthy research direction. Second, MIL techniques can be incorporated into CBIR systems to deal with

the ambiguity existing in the user queries. One of the key problems in developing a practical multi-instance learning based CBIR system is to obtain a nice bag generator. Thus how to formulate an appropriate bag generator for MIL problems should be paid special attention to. Third, how to efficiently integrate MIL with other methods based on the tradeoff between computational complexity and model reconstruction error is a valuable research direction in the future. Fourth, due to the multi-instance representation allows for concept descriptions that are defined upon the interaction of instance-level concepts, which is a natural way to describe visual concepts. As a consequence how to find more effective and generally applicable algorithms for learning visual concepts is also a promising research direction. Fifth, since labeled images are often hard to obtain or create in large quantities in practical applications while the unlabeled ones are easier to collect from the image repository. Hence, semi-supervised learning method, which aims at learning from labeled and unlabeled data simultaneously, can be employed to boost the quality of the training image data with the help of unlabelled data in the presence of the small sample size problem. Particularly, how to solve MIL issues for partially labeled data has become a promising research direction to leverage informative yet unlabeled data. Last but not the least, for the future work, MIL should be applied in more wider ranges to deal with more multimedia related tasks, such as speech recognition, action recognition, music information retrieval and other multimedia event detection tasks, etc.

Acknowledgment. The author would like to sincerely thank the anonymous reviewers for their valuable comments and insightful suggestions that have helped to improve the paper. In addition, the author thanks Prof. Zhongzhi Shi for stimulating discussions and helpful hints. This work is partially supported by the National Program on Key Basic Research Project (No.2013CB329502), the National Natural Science Foundation of China (No.61202212), the Key R&D Program of the Shaanxi Province of China (No.2018GY-037) and the Special Research Project of the Educational Department of Shaanxi Province of China (No.18JK0051).

REFERENCES

[1] P. Duygulu, K. Barnard, J. De Freitas, et al., Object recognition as machine translation: learning a lexicon for a fixed image vocabulary, *Proc. of the 7th European Conf. on Computer Vision (ECCV'02)*, pp. 97–112, 2002.

[2] L. Jeon, V. Lavrenko and R. Manmantha, Automatic image annotation and retrieval using cross-media relevance model, *Proc. of the 26th Int'l Conf. on Research and Development in Information Retrieval (SIGIR'03)*, pp. 119–126, 2003.

[3] R. Manmatha, V. Lavrenko and J. Jeon, A model for learning the semantics of pictures, *Advances in Neural Information Processing Systems 16 (NIPS'03)*, pp. 553–560, 2003.

[4] S. Feng, R. Manmatha and V. Lavrenko, Multiple Bernoulli relevance models for image and video annotation, *Proc. of the Int'l Conf. on Computer Vision and Pattern Recognition (CVPR'04)*, pp. 1002–1009, 2004.

[5] J. Liu, B. Wang, M. Li, et al., Dual cross-media relevance model for image annotation, *Proc. of the 15th Int'l Conf. on Multimedia (MM'07)*, pp. 605–614, 2007.

[6] F. Monay and D. Gatica-Perez, On image auto-annotation with latent space models, *Proc. of the 11th Int'l Conf. on Multimedia (MM'03)*, pp. 275–278, 2003.

[7] F. Monay and D. Gatica-Perez, PLSA-based image auto-annotation: constraining the latent space, *Proc. of the 12th Int'l Conf. on Multimedia (MM'04)*, pp. 348–351, 2004.

[8] F. Monay and D. Gatica-Perez, Modeling semantic aspects for cross-media image indexing, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 10, pp. 1802–1817, 2007.

[9] D. Tian, X. Zhao and Z. Shi, An efficient refining image annotation technique by combining probabilistic latent semantic analysis and random walk model, *Intelligent Automation & Soft Computing*, vol. 20, no. 3, pp. 335–345, 2014.

[10] D. Tian, W. Zhang, X. Zhao, et al., Employing PLSA model and max-bisection for refining image annotation, *Proc. of the 20th Int'l Conf. on Image Processing (ICIP'13)*, pp. 3996–4000, 2013.
[11] H. Xu, J. Wang, X. Hua, et al., Tag refinement by regularized LDA, *Proc. of the 17th Int'l Conf. on Multimedia (MM'09)*, pp. 573–576, 2009.
[12] A. Makadia, V. Pavlovic and S. Kumar, A new baseline for image annotation, *Proc. of the 13th European Conf. on Computer Vision (ECCV'08)*, pp. 316–329, 2008.
[13] M. Guillaumin, T. Mensink, J. Verbeek, et al., TagProp: discriminative metric learning in nearest neighbor models for image auto-annotation, *Proc. of the 12th Int'l Conf. on Computer Vision (ICCV'09)*, pp. 309–316, 2009.
[14] O. Maron and T. Lozano-Perez, A framework for multiple-instance learning, *Advances in Neural Information Processing Systems 11 (NIPS'98)*, pp. 570–576, 1998.
[15] Z. Zhou and M. Zhang, Multi-instance multi-label learning with application to scene classification, *Advances in Neural Information Processing Systems 20 (NIPS'07)*, pp. 1609–1616, 2007.
[16] S. Andrews, I. Tsochantaridis and T. Hofmann, Support vector machines for multiple-instance learning, *Advances in Neural Information Processing Systems 15 (NIPS'02)*, pp. 561–568, 2002.
[17] S. Andrews, T. Hofmann and I. Tsochantaridis, Multiple instance learning with generalized support vector machines, *Proc. of the 18th National Conf. on Artificial Intelligence (AAAI'02)*, pp. 943–944, 2002.
[18] P. Gehler and O. Chapelle, Deterministic annealing for multiple-instance learning, *Proc. of the 11th Int'l Conf. on Artificial Intelligence and Statistics (AISTATS'07)*, pp. 123–130, 2007.
[19] C. Yang, M. Dong and F. Fotouhi, Region-based image annotation through multiple-instance learning, *Proc. of the 13th Int'l Conf. on Multimedia (MM'05)*, pp. 435–438, 2005.
[20] C. Yang, M. Dong and J. Hua, Region-based image annotation using asymmetrical support vector machine-based multiple-instance learning, *Proc. of the Int'l Conf. on Computer Vision and Pattern Recognition (CVPR'06)*, pp. 2057–2063, 2006.
[21] S. Gao and Q. Sun, Exploiting generalized discriminative multiple instance learning for multimedia semantic concept detection, *Pattern Recognition*, vol. 41, no. 10, pp. 3214–3223, 2008.
[22] Y. Zhao, Y. Zhao, Z. Zhu, et al., MRS-MIL: minimum reference set based multiple instance learning for automatic image annotation, *Proc. of the 15th Int'l Conf. on Image Processing (ICIP'08)*, pp. 2160–2163, 2008.
[23] S. Feng, D. Xu and B. Li, Automatic region-based image annotation using an improved multiple-instance learning algorithm, *Chinese Journal of Electronics*, vol. 17, no. 1, pp. 43–47, 2008.
[24] S. Feng and D. Xu, Transductive multi-instance multi-label learning algorithm with application to automatic image annotation, *Expert Systems with Applications*, vol. 37, no. 1, pp. 661–670, 2010.
[25] R. Rahmani and S. Goldman, MISSL: multiple-instance semi-supervised learning, *Proc. of the 23rd Int'l Conf. on Machine Learning (ICML'06)*, pp. 705–712, 2006.
[26] X. Qi and Y. Han, Incorporating multiple SVMs for automatic image annotation, *Pattern Recognition*, vol. 40, no. 4, pp. 728–741, 2007.
[27] X. Xue, W. Zhang, J. Zhang, et al., Correlative multi-label multi-instance image annotation, *Proc. of the 13th Int'l Conf. on Computer Vision (ICCV'11)*, pp. 651–658, 2011.
[28] D. Wang, J. Li and B. Zhang, Multiple-instance learning via random walk, *Proc. of the 17th European Conf. on Machine Learning (ECML'06)*, pp. 473–484, 2006.
[29] S. Zhu and X. Tan, A novel automatic image annotation method based on multi-instance learning, *Procedia Engineering*, vol. 15, pp. 3439–3444, 2011.
[30] C. Nguyen, D. Zhan and Z. Zhou, Multi-modal image annotation with multi-instance multi-label LDA, *Proc. of the 23rd Int'l Joint Conf. on Artificial Intelligence (IJCAI'13)*, pp. 1558–1564, 2013.
[31] Z. Xia, J. Peng, X. Feng, et al., Multiple instance learning for automatic image annotation, *Proc. of the 19th Int'l Conf. on Multimedia Modeling (MMM'13)*, Part II, LNCS 7733, pp. 194–205, 2013.
[32] O. Maron and A. Ratan, Multiple-instance learning for natural scene classification, *Proc. of the 15th Int'l Conf. on Machine Learning (ICML'98)*, pp. 341–349, 1998.
[33] C. Yang and T. Lozano-Perez, Image database retrieval with multiple-instance learning techniques, *Proc. of the 16th Int'l Conf. on Data Engineering (ICDE'00)*, pp. 233–243, 2000.
[34] Z. Zhou, M. Zhang and K. Chen, A novel bag generator for image database retrieval with multi-instance learning techniques, *Proc. of the 15th Int'l Conf. on Tools with Artificial Intelligence (ICTAI'03)*, pp. 565–569, 2003.
[35] Q. Zhang, S. Goldman, W. Yu, et al., Content-based image retrieval using multiple-instance learning, *Proc. of the 19th Int'l Conf. on Machine Learning (ICML'02)*, pp. 682–689, 2002.

[36] C. Zhang and X. Chen, Region-based image clustering and retrieval using multiple instance learning, *Proc. of the Int'l Conf. on Image and Video Retrieval (CIVR'05)*, pp. 194–204, 2005.
[37] X. Yuan, X. Hua, M. Wang, et al., A novel multiple instance learning approach for image retrieval based on Adaboost feature selection, *Proc. of the Int'l Conf. on Multimedia and Expo (ICME'07)*, pp. 1491–1494, 2007.
[38] D. Zhang, Z. Shi, Y. Song, et al., Localized content-based image retrieval using semi-supervised multiple instance learning, *Proc. of the 8th Asian Conf. on Computer Vision (ACCV'07)*, pp. 180–188, 2007.
[39] C. Wang, L. Zhang and H. Zhang, Graph-based multiple-instance learning for object-based image retrieval, *Proc. of the Int'l Conf. on Multimedia Information Retrieval (MIR'08)*, pp. 156–163, 2008.
[40] Q. Qiao and P. Beling, Localized content based image retrieval with self-taught multiple instance learning, *Proc. of the Int'l Conf. on Data Mining Workshops (ICDMW'09)*, pp. 170–175, 2009.
[41] W. Li, L. Duan, D. Xu, et al., Text-based image retrieval using progressive multi-instance learning, *Proc. of the 13th Int'l Conf. on Computer Vision (ICCV'11)*, pp. 2049–2055, 2011.
[42] D. Li, J. Peng, Z. Li, et al., LSA based multi-instance learning algorithm for image retrieval, *Signal Processing*, vol. 91, no. 8, pp. 1993–2000, 2011.
[43] D. Li, J. Fan, D. Wang, et al., Latent topic based multi-instance learning method for localized content-based image retrieval, *Computers and Mathematics with Applications*, vol. 64, no. 4, pp. 500–510, 2012.
[44] M. Kim and F. Torre, Multiple instance learning via Gaussian processes, *Data Mining and Knowledge Discovery*, vol. 28, no. 4, pp. 1078–1106, 2014.
[45] Y. Xu, Multiple-instance learning based decision neural networks for image retrieval and classification, *Neurocomputing*, vol. 171, pp. 826–836, 2016.
[46] X. Wang, D. Wei, H. Cheng, et al., Multi-instance learning based on representative instance and feature mapping, *Neurocomputing*, vol. 216, pp. 790–796, 2016.
[47] Y. Chen and J. Wang, Image categorization by learning and reasoning with regions, *Journal of Machine Learning Research*, vol. 5, no. 8, pp. 913–939, 2004.
[48] Y. Chen, J. Bi and J. Wang, MILES: multiple-instance learning via embedded instance selection, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 12, pp. 1931–1947, 2006.
[49] H. Pao, S. Chuang, Y. Xu, et al., An EM based multiple instance learning method for image classification, *Expert Systems with Applications*, vol. 35, no. 3, pp. 1468–1472, 2008.
[50] J. Li and J. Li, A novel semi-supervised multi-instance learning approach for scene recognition, *Proc. of the 9th Int'l Conf. on Fuzzy Systems and Knowledge Discovery (FSKD'12)*, pp. 1206–1210, 2012.
[51] X. Xi, X. Xu and X. Wang, A novel multi-instance learning algorithm with application to image classification, *Proc. of the Asia-Pacific Signal & Information Processing Association Annual Summit and Conference (APSIPA ASC'12)*, pp. 1–6, 2012.
[52] X. Song, L. Jiao, S. Yang, et al., Sparse coding and classifier ensemble based multi-instance learning for image categorization, *Signal Processing*, vol. 93, no. 1, pp. 1–11, 2013.
[53] D. Li, J. Wang, X. Zhao, et al., Multiple kernel-based multi-instance learning algorithm for image classification, *Journal of Visual Communication and Image Representation*, vol. 25, no. 5, pp. 1112–1117, 2014.
[54] Z. Qi, Y. Tian, X. Yu, et al., A multi-instance learning algorithm based on nonparallel classifier, *Applied Mathematics and Computation*, vol. 241, pp. 233–241, 2014.
[55] S. Goldman and S. Scott, Multiple-instance learning of real-valued geometric patterns, *Annals of Mathematics and Artificial Intelligence*, vol. 39, no. 3, pp. 259–290, 2003.
[56] J. He, H. Gu and Z. Wang, Multi-instance multi-label learning based on Gaussian process with application to visual mobile robot navigation, *Information Sciences*, vol. 190, pp. 162–177, 2012.
[57] G. Jia, Y. Feng and B. Zheng, Medical image semantic annotation based on MIL, *Proc. of the Int'l Conf. on Complex Medical Engineering (ICME'13)*, pp. 85–90, 2013.
[58] Y. Xu, T. Mo, Q. Feng, et al., Deep learning of feature representation with multiple instance learning for medical image analysis, *Proc. of the 39th Int'l Conf. on Acoustics, Speech and Signal Processing (ICASSP'14)*, pp. 1626–1630, 2014.
[59] J. Melendez, B. Ginneken, P. Maduskar, et al., On combining multiple-instance learning and active learning for computer-aided detection of tuberculosis, *IEEE Transactions on Medical Imaging*, vol. 35, no. 4, pp. 1013–1024, 2016.

[60] M. Yousefi, A. Krzyak and C. Suen, Mass detection in digital breast tomosynthesis data using convolutional neural networks and multiple instance learning, *Computers in Biology and Medicine*, vol. 96, pp. 283–293, 2018.
[61] H. Liu, C. Zhou, J. Shen, et al., Video caption detection algorithm based on multiple instance learning, *Proc. of the 5th Int'l Conf. on Internet Computing for Science and Engineering (ICICSE'10)*, pp. 20–24, 2010.
[62] X. Xu, Y. Jiang, X. Xue, et al., Semi-supervised multi-instance multi-label learning for video annotation task, *Proc. of the 20th Int'l Conf. on Multimedia (MM'12)*, pp. 737–740, 2012.
[63] R. Bunescu and R. Mooney, Learning to extract relations from the web using minimal supervision, *Proc. of the 45th Annual Meeting of the Association for Computational Linguistics (ACL'07)*, pp. 576–583, 2007.
[64] A. Zafra, E. Gibaja and S. Ventura, Multiple instance learning with multiple objective genetic programming for web mining, *Applied Soft Computing*, vol. 11, no. 1, pp. 93–102, 2011.
[65] D. Tarrag, C. Cornelis, R. Bello, et al., A multi-instance learning wrapper based on the Rocchio classifier for web index recommendation, *Knowledge-Based Systems*, vol. 59, pp. 173–181, 2014.
[66] P. Viola, J. Platt and C. Zhang, Multiple instance boosting for object detection, *Advances in Neural Information Processing Systems 19 (NIPS'06)*, pp. 1419–1426, 2006.
[67] S. Phan, D. Le and S. Satoh, Multimedia event detection using event-driven multiple instance learning, *Proc. of the 23rd Int'l Conf. on Multimedia (MM'15)*, pp. 1255–1258, 2015.
[68] R. Cinbis, J. Verbeek and C. Schmid, Weakly supervised object localization with multi-fold multiple instance learning, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 1, pp. 189–203, 2017.
[69] C. Xu, W. Tao, Z. Meng, et al., Robust visual tracking via online multiple instance learning with fisher information, *Pattern Recognition*, vol. 48, no. 12, pp. 3917–3926, 2015.
[70] M. Abdechiri, K. Faez and H. Amindavar, Visual object tracking with online weighted chaotic multiple instance learning, *Neurocomputing*, vol. 247, pp. 16–30, 2017.
[71] D. Li, G. Wen, Y. Kuai, et al., Spatio-temporally weighted multiple instance learning for visual tracking, *Optik*, vol. 171, pp. 904–917, 2018.
[72] M. Dundar, G. Fung, B. Krishnapuram, et al., Multiple-instance learning algorithms for computer-aided detection, *IEEE Transactions on Biomedical Engineering*, vol. 55, no. 3, pp. 1015–1021, 2008.
[73] I. Gondra and T. Xu, A multiple instance learning based framework for semantic image segmentation, *Multimedia Tools and Applications*, vol. 48, no. 2, pp. 339–365, 2010.
[74] Y. Shen and J. Fan, Leveraging loosely-tagged images and inter-object correlations for tag recommendation, *Proc. of the 18th Int'l Conf. on Multimedia (MM'10)*, pp. 5–14, 2010.
[75] D. Li, N. Li, J. Wang, et al., Pornographic images recognition based on spatial pyramid partition and multi-instance ensemble learning, *Knowledge-Based Systems*, vol. 84, pp. 214–223, 2015.
[76] S. Bandyopadhyay, D. Ghosh, R. Mitra, et al., MBSTAR: multiple instance learning for predicting specific functional binding sites in microRNA targets, *Scientific Reports*, vol. 5, id. 8004, 2015.
[77] J. Stiborek, T. Pevn and M. Rehk, Multiple instance learning for malware classification, *Expert Systems with Applications*, vol. 93, pp. 346–357, 2018.
[78] J. Amores, Multiple instance classification: review, taxonomy and comparative study, *Artificial Intelligence*, vol. 201, pp. 81–105, 2013.
[79] M. Carbonneau, V. Cheplygina, E. Granger, et al., Multiple instance learning: A survey of problem characteristics and applications, *Pattern Recognition*, vol. 77, pp. 329–353, 2018.
[80] F. Herrera, S. Ventura, R. Bello, et al., Multiple Instance Learning: Foundation and Algorithms, Springer, 2016.