Multi-Instance Classification by Max-Margin Training of Cardinality-Based Markov Networks

Hossein Hajimirsadeghi and Greg Mori

Abstract—We propose a probabilistic graphical framework for multi-instance learning (MIL) based on Markov networks. This framework can deal with different levels of labeling ambiguity (i.e., the portion of positive instances in a bag) in weakly supervised data by parameterizing cardinality potential functions. Consequently, it can be used to encode different cardinality-based multi-instance assumptions, ranging from the standard MIL assumption to more general assumptions. In addition, this framework can be efficiently used for both binary and multiclass classification. To this end, an efficient inference algorithm and a discriminative latent max-margin learning algorithm are introduced to train and test the proposed multi-instance Markov network models. We evaluate the performance of the proposed framework on binary and multi-class MIL benchmark datasets as well as two challenging computer vision tasks: cyclist helmet recognition and human group activity recognition. Experimental results verify that encoding the degree of ambiguity in data can improve classification performance.

Index Terms—Multiple Instance Learning, Markov Network, Conditional Random Field, Cardinality Models.

1 Introduction

Multi-instance learning (MIL) aims to recognize patterns from weakly supervised data. Contrary to standard supervised learning, where each training instance is labeled, in the MIL paradigm a *bag of instances* share a label. For example in the binary MIL, each bag of instances is labeled as positive or negative. The training data is given as labeled bags, and the goal is to predict the label of test bags. In the standard binary multi-instance (MI) assumption, a bag is positive if it contains *at least one* positive instance, while in a negative bag all the instances are negative. This ambiguity in the instance labels is passed on to the learning algorithm, which should incorporate the information to classify unseen bags. In this work we develop a novel framework for MIL, which can model more general multi-instance assumptions and deal with different levels of labeling ambiguity in the bags.

The standard MI assumption (i.e., at least one instance in a positive bag is positive) is a too weak assumption in many MIL applications. For example, in the cyclist helmet recognition problem shown in Fig. 1, the goal is to detect if the cyclist in the video is wearing helmet, given the automatically estimated track of the cyclist's head position. This can be modeled as a MIL problem, where the cyclist track is represented as a bag of image patches extracted around the estimated cyclist's head position in each frame. Because of the imperfect tracking, not all extracted windows are centered on the helmet, and consequently not all instances in a positive bag are positive. However, the positive instances are not sparse in the positive bags, either. In fact, many instances are true positives and not just irrelevant elements in a bag. Using this prior information can help to train stronger classifiers. Further, because of noisy, occluded, or low-quality feature representations, negative bags can also contain instances that are effectively indistinguishable from positive instances. In these cases more robust MI assumptions are needed, and this paper contributes in this direction.

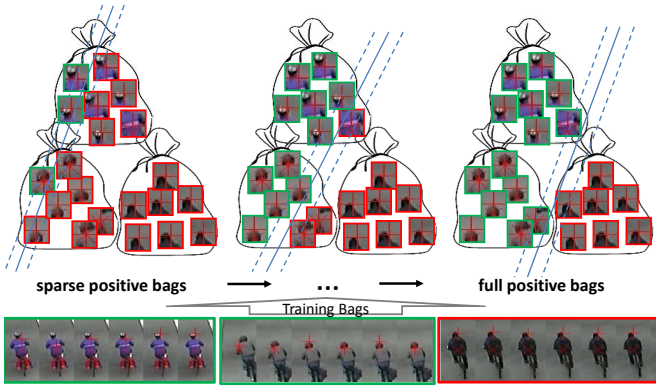


Fig. 1. Cyclist helmet recognition using the proposed max-margin MIL method. The goal is to recognize if the cyclist is wearing a helmet or not, given the input video. Each video is treated as a bag of instances, where each instance is represented by an automatically detected window around the cyclist's head. The proposed cardinality-based models help to control the positive/negative label proportions in the bags and encode a wide range of multi-instance assumptions.

On the other hand, analysis of the cardinality-based relations is intuitive and intrinsic to some visual recognition tasks. For example, in collective activity recognition (e.g. [1]) the primary approach to analyze the activity of a group of people is to look at the actions of individuals in a scene. There have been various methods for modeling the *structure* of a group activity [2, 3, 4], capturing spatio-temporal relations between people in a scene. However, these methods do not directly consider cardinality relations about the *number* of people that should be involved in an activity. These cardinality relations vary per activity. An activity like falling

H. Hajimirsadeghi and G. Mori are with School of Computing Science, Simon Fraser University, Burnaby, BC, CANADA e-mails: hosseinh@sfu.ca, mori@cs.sfu.ca
Manuscript received

in a nursing home [2] is different in composition from an activity such as queuing [3], involving different numbers of people (one person falls, many people queue). Further, noise and clutter, in the form of people in a scene performing irrelevant actions, confounds recognition algorithms.

To address these issues, we develop a general MIL framework to encode various types of cardinality relations and make a flexible notion of labeled bags. This framework is built on a latent structured model based on Markov networks to incorporate count-based measurements over instances, which can extend from the notion of "at least one positive" to "at least some positives" to "nearly all positives." Thus, it can (1) deal with different levels of ambiguity or clutter in the data and (2) encode various kinds of cardinality relations/constraints on instances, either predefined by the user or learned directly from the data. Indeed, this framework can be even adapted to estimate the appropriate MIL notion from training data without prior assumption on the proportion of positives in the bags.

A preliminary version of this work was published previously in [5]. This paper extends on this work, adding algorithms for *multi-class* multi-instance classification, introducing new applications and additional empirical evaluation. In sum, this paper presents the following contributions. First, we show that the proposed framework can be used for multi-class MIL without converting the problem to multiple binary classification (e.g., employing exhustive one-vs-all or one-vs-one appraoches), commonly used in MIL methods. Second, it is shown that the proposed Markov network facilitates modeling of the inter-relations between different components of a bag. It helps to integrate the local information of the instances (i.e., instance-level information) with the global information elicited from the whole bag (i.e., bag-level information). For example, an image can be represented by local feature vectors extracted from several regions of interest in the image as well as a global feature vector extracted from the whole image. Third, we propose exact and efficient inference algorithms to evaluate these general MIL models efficiently without any approximation. For the learning criterion, we propose a latent max-margin discriminative algorithm to train the models.

This paper is organized as follows. Section 2 reviews related work and provides a qualitative comparison between this work and the previous works. Section 3 describes our framework of MIL with Markov networks. In particular, the models for different MI assumptions, including the standard MI assumption and more general MI assumptions are described in this section. In Section 4 the inference and learning algorithms are explained. Section 5 presents the experimental studies on MIL benchmark datasets as well as cyclist helmet classification and human group activity recognition problems. We conclude in Section 6.

2 Related Work

MIL methods can be categorized based on different criteria such as the learning approach (e.g. maximum likelihood, max-margin, etc.), the multi-instance assumption (e.g. standard assumption, ratio-based assumption, etc.) [6], or the space/level that the discriminative information lies in the method (instance-space vs. bag-space) [7]. In this section, we review a variety of MIL methods in two subsections of instance-space methods and bag-space methods. However, we also try to briefly explain the learning approach and the multi-instance assumption used in each method.

2.1 Instance-Space Methods

Instance-level methods classify bags by aggregation of instance-level classification scores. To this end, an instance-level classifier is trained to classify positive and negative instances in the instance space, and based on these classifiers a bag-level classifier is obtained.

2.1.1 Methods Encoding Standard MI Assumption

Dietterich et al. [8] introduced the early algorithms for multi-instance learning. The main idea was to construct a hyper-rectangle maximizing the number of bags with at least one instance in that rectangle while excluding all instances of negative bags. So, this algorithm encodes the standard MI assumption. Based on similar ideas, the diverse density (DD) framework [9] was proposed for MIL. This approach works by finding a *concept point* which is near to at least one instance of every positive bag, but far from all negative instances (i.e., standard MI assumption). Finding this point is formulated as maximizing the diverse density function, which is in fact the likelihood function of training bags. EM-DD [10] is the expectation-maximization (EM) version of DD, which incorporates the iterative EM approach of estimating positive instances and updating the concept hypothesis within the DD framework.

Andrews et al. [11] modified SVMs for MIL by proposing two max-margin algorithms. The first, mi-SVM, aims to maximize the instance margin jointly over the hidden instance labels. The second, MI-SVM, tries to maximize the bag margin, where the bag margin is defined by the most positive instance of each bag (a.k.a witness instance). Both these algorithms are formulated as mixed-integer optimization problems, which are solved approximately by iterating over two steps of inferring the instance labels (in mi-SVM) or finding the witness instance (in MI-SVM) and then continuous optimization of the SVM weight vectors using the instances. Following the same approach, Mangasarian and Wild [12] proposed MICA. MICA is an extension of MI-SVM, which does not explicitly identify a specific witness instance in a bag but finds a convex combination of the instances as a witness. Bunescu and Mooney [13] used the transductive SVM framework to propose a modified version of mi-SVM which can more directly enforce the standard MI assumption and perform more effectively for sparse positive bags. AL-SVM and AW-SVM [14] are the other extensions of mi-SVM and MI-SVM, which apply deterministic annealing to the mixed-integer programs of the multi-instance SVM formulations in order to find more accurate solutions. Later, this idea was used in MI-Forests [15] to perform the mixed-integer optimization of a margin-dependent loss function over randomized trees, using deterministic annealing.

The very successful Latent SVM [16] is also a max-margin MIL method. For positive instances, a set of latent variable values is used. One can consider the set of completed data instances (latent variable values with observed input feature values) as a *bag* in MIL, similar to the MI-SVM framework. Latent SVM has been used in numerous

applications, and often obtain very successful performance. However, it uses the "at least one positive instance" assumption for positive bags. As noted above, for some applications this is limiting since many latent variable settings could in fact be positive and could aid in training a better classifier. The more general MI assumptions and algorithms in this paper aim to remedy this.

2.1.2 Methods Encoding Non-Standard MI Assumptions

In recent years, more general MIL algorithms have been developed to address non-standard multi-instance assumptions such as *ratio-based* assumptions [17, 18], where the proportion of positive instances in a bag determines the bag label. Gehler and Chapelle [17] proposed ALP-SVM, which can control the expected ratio of positive instances in the bags. They argued that different levels of ambiguity in positive bags can influence the performance of MIL methods. Hence, they provided the possibility to encode prior knowledge about the data set, i.e., fraction of positive instances (witnesses) in a positive bag. This algorithm need a preset parameter which determines the fixed ratio of witnesses.

Li et al. [18] proposed MIL-CPB, an algorithm for multi-instance learning with constrained positive bags. This model uses a generalized MI assumption, where the positive bags contain at least a certain portion of positive instances (i.e., ratio-constrained assumption). The formulation of MIL-CPB is similar to the mixed-integer formulation of mi-SVM but with more general constraints on instance labels. It is shown that this NP-hard problem can be viewed as a multiple kernel learning problem with an exponential number of base kernels. Solving this problem is intractable in practice. Li et al. proposed an iterative cutting-plane algorithm to find a subset of feasible solutions which can adequately approximate the original problem.

Hajimirsadeghi and Mori [19] proposed MIRealBoost, a boosting framework for MIL which can softly explore different levels of ambiguity using linguistic aggregation functions with different degrees of orness. Hence, the notion of positive bag is extended to a wider and more intuitive range of assumptions. This algorithm also needs approximate before-hand knowledge of ambiguity level (e.g. the witness ratio), or should use cross-validation to estimate it. Yu et al. [20] proposed $\propto$SVM for learning from instance label proportions. This SVM-based model also tries to control the ratio of positive instances in a bag. They proposed two algorithms to learn the model: (1) alternating optimization of the mixed-integer programming problem and (2) convex relaxation of the objective function.

Despite successful results of the algorithms above, almost all of them use some kind of heuristics or relaxation and consequently provide approximate solutions to the general problem of multi-instance learning based on label proportions or lack solid mathematical proof of convergence. In addition, they are limited to specific cardinality assumptions (e.g. ratio-constrained assumptions) and to capture new cardinality relations between the instance labels the proposed models or learning algorithms should be modified.

2.1.3 Methods based on Probabilistic Graphical Models

Probabilistic graphical models (PGMs) are powerful tools to capture inter-relations between random variables and learn structured models. Thus, they can be assumed a natural fit to model multi-instance problems. Note that although we have categorized PGM-based methods as instance-space methods, these methods fall close to the boundaries of bag-space methods. In fact, in PGMs both instance-level local information and bag-level global information can be modeled and mixed. However, since the base of these models is built on the instances, and the first-level discrimination lies in the instance space, we think that PGM-based MIL methods are mostly (not always) closer to instance-space methods.

Warrell and Torr [21] developed an algorithm for MIL based on structured bag models. This method constructs a conditional random filed (CRF) with energy functions defined on the instances, instance labels and the bag label. In this model, hard and soft constraints are presented on the instance labels to encode standard MI assumption as well as more general soft ratio-based assumptions. Given the proposed CRF and the constraints, the instance and bag labels are inferred *approximately* by dual decomposition, and the models are trained by likelihood maximization using deterministic annealing.

Deselaers and Ferrari [22] proposed MI-CRF. In this method, the bags are modelled as nodes in a CRF, where each node can take one of the instances of the bag as its state. So, the bags are jointly trained and classified in this model. Louradour and Larochelle [23] proposed extensions of restricted Boltzmann machines (RBMs) for classifying sets of instances. In the proposed method, RBMs are extended by duplicating the visible and hidden layers for each instance. The basic idea is to encode the bag label besides the input instance vectors in the visible layer and embedding the constraints in the hidden layer. Adel et al. [24] proposed a general framework to use generative graphical models in the MIL paradigm. This framework studies and analyzes different Bayes net structures for MIL.

2.2 Bag-Space Methods

Bag-space methods treat each bag as a whole entity and train a classifier directly on the bags by making a global representation of bags or extracting discriminative bag-level information from them. In this section, we briefly explain these methods, classified in three main subcategories: "embedded-space" methods, "kernel-based" methods, and "distance-based" methods.

2.2.1 Embedded-Space Methods

The methods described in this section transform MIL problem to a standard classification problem by mapping the bags into an embedded single-instance space. Simple MI [25] is a very simple and fast algorithm of this type. Each bag is mapped to the average of its instances. The averaging can be performed by arithmetic mean or geometric mean. Although this algorithm is very simple, surprisingly, it has shown successful results in some MIL problems (e.g., when the positive bags have mostly positive instances – i.e., less instance label ambiguity). Another family of embedded-space methods are Histogram-Based Methods [7], which

work similar to bag-of-words (BOW) methods by mapping each bag to a histogram vector, using a vocabulary. First, a vocabulary of concepts (or words) is obtained by hard or soft clustering of all instances in the training bags. Next, each bag is mapped to a histogram vector of the concepts.

DD-SVM [26] and MILES [14] are two algorithms, which combine the diverse density approach with SVM classification. Both these algorithms use the concept points introduced in the diverse density framework to convert each bag to a new single-instance feature vector. Next, a standard L_2-norm SVM classifier is trained in DD-SVM. However, in MILES, an L_1-norm SVM is used. The L_1-norm SVM can be employed for both classification and concept point selection.

2.2.2 Kernel-Based Methods

Kernel-Based Methods work by defining kernels on the bags. As a result, any standard kernel machine can be used for classification. Note that kernel-based methods also works by performing an implicit space transformation and mapping. Thus, it might be also possible to categorize kernel-based methods as embedded-space methods.

Gartner et al. [27] introduced a class of multi-instance kernels (MI-Kernels), which are variants of set kernels [28]. The standard MI-kernel is a bag-level kernel which is obtained by summing up instance-level kernels on all instance pairs of two bags. The proposed MI-kernel assumes equal weights on all instances of a bag. However, usually in positive bags all the instances are not equally important. To alleviate this problem, later, Kwok and Cheung [29] proposed marginalized MI kernels. These kernels specify the importance of an instance pair from two bags according to the consistency of their probabilistic instance labels.

Zhou et al. [30] proposed two graph-based algorithms, MIGraph and miGraph, for multi-instance learning. Both algorithms work by mapping a bag into an undirected graph and then designing a graph kernel. MIGraph constructs a weighted ϵ-graph for every bag. In this graph, each instance is modeled as a node, and every two nodes are connected if the Euclidean distance between the two instances is less than a preset threshold ϵ. Next, a kernel function is defined between bags by aggregating the base kernels on node pairs and edge pairs. MIGraph has high computational complexity due to the large number of edges usually existing in the constructed graph. But, miGraph is more computationally efficient. miGraph implicitly maps a bag to a graph by only creating the affinity matrix of the graph. Given this affinity matrix, a bag-level kernel is defined which is independent of the number of edges.

2.2.3 Distance-Based Methods

A class of MIL algorithms uses distance metrics to classify bags. The distance can be a bag-to-bag (B2B) distance or a class-to-bag (C2B) distance. Also, the distance metric can be fixed or learned from training data. For example, Citation kNN [32] applies a B2B distance in a generalized and more robust k Nearest Neighbor (kNN) algorithm.

Wang et al. [33] proposed an algorithm to learn a robust and discriminative C2B distance for MIL. Unlike the multi-instance distances defined in the similar previous works (e.g., [34, 35, 36]), the proposed distance is based on not-squared l_2-norm distance. It is well-known that not-squared l_2-norm distance is robust againts outliers [37], which makes it suitable for MI data, where the outlier instances abound because of label ambiguity in positive bags. Learning the distance function is formulated as minimizing the C2B distance from a class to all its bags, while maximizing the distance to all bags of other classes.

2.3 Our Work

In this work, we propose a MIL framework based on Markov networks (which is a PGM). This framework uses cardinality potentials [38, 39] to model general MI assumptions, and superior to the similar previous works [17, 40, 18, 19, 21, 20], which follow nonstandard MI assumptions, it presents the following contributions. First, it can encode any cardinality-based multi-instance assumption[1]. It can even work without prior assumption on the cardinality of positive instances inside the bags and be trained to discover this knowledge directly from data. Second, it can be used for both binary and multi-class classification. Third, the inference and learning of the proposed models is exact and no approximation or heuristics are required. Finally, the proposed model allows flexible integration of bag-level and instance-level information in a bag, leveraging benefits from both global and local representations of the bag in both bag and instance spaces.

To conclude this section, Tabel 1 provides a summary of the algorithms reviewed above. In this table, the MI assumption followed in each method is also given. The ratio-based assumption refers to any assumption which is based on the instance label proportions in a bag. Ratio-constrained assumption is a special ratio-based assumption, which is an immediate extension of the standard MI assumption and assumes a bag is positive if at least a certain ratio of the instances are positive. Metadata assumption is used by convention to refer to the assumption used in embedded-space and kernel-based methods [6]. This assumption originates from the fact that in these methods classification is performed in a metadata embedding space.

3 MIL Using Markov Networks

In MIL, training examples are presented in bags where the instances in a bag share a label. In this work, we use Markov networks to model MIL problems and develop a generalized notion of labeled bags. The proposed Markov networks are used to define a scoring function for bag classification.

3.1 The Proposed Markov Network for MIL

In this section, we firstly introduce the model for binary multi-instance classification and next extend it for multiclass classification.

3.1.1 Binary Classification

Let $\mathcal{B} = \{\mathcal{I}_1, \cdots, \mathcal{I}_m\}$ denote a bag with m instances and a binary bag label $Y \in \{-1, 1\}$. Each instance $\mathcal{I}_i$ is represented by a fixed-length feature vector $\mathbf{x}_i = [x_{i1}, \cdots, x_{iD}] \in$

1. Although we focus on ratio-based cardinality assumptions in this work, the proposed model is not limited to these assumptions and can encode any cardinality-based assumptions on the instance labels.

TABLE 1
A list of some well-known MIL methods

Method	Summary of the algorithm	Base discrimination space/level	Multi-instance assumption
Axis-Parallel Rectangles [8]	Finding a hyper-rectangle that maximizes the number of positive bags which have at least one instance in this region, but excludes instances of negative bags as much as possible.	Instance space	Standard assumption
Diverse Density [9]	Estimating the probability of instances based on distance from an instance prototype, which is close to at least one instance of every positive training bag but far from instnaces of all negative training bags.	Instance space	Standard assumption
EM-DD [10]	Using expectation-maximization to maximize the diverse density function.	Instance space	Standard assumption
mi-SVM [11]	Maximizing the instance margin jointly over the latent instance labels, using an iterative algorithm.	Instance space	Standard assumption
MI-SVM [11]	Maximizing the bag margin in an iterative procedure, where at each iteration every positive bag is represented by the most postive instance of the bag.	Instance space	Standard assumption
sMIL, stMIL [13]	sMIL modifies miSVM constraints to be more effective for sparse positive bags. stMIL is the transductive SVM version of sMIL.	Instance space	Standard assumption
AL-SVM, AW-SVM, ALP-SVM [17]	Optimizing mi-SVM and MI-SVM objective functions with deterministic annealing.	Instance space	Standard assumption for AL-SVM & AW-SVM. Ratio-based for ALP-SVM.
MI-Forests [15]	Optimizing a confidence maximizing loss function over randomized trees, using an iterative DA-based method.	Instance space	Standard assumption
MILBoost [31]	Maximizing the log likelihood of training bags using AnyBoost framework.	Instance space	Standard assumption
MIL-CPB [18]	Optimizing SVM-like objective functions with ratio-based MIL constraints for the positive bags, using an iterative cutting plane algorithm.	Instance space	Ratio-constrained assumption
MIRealBoost [19]	Maximizing the expected log likelihood of training bags, using standard RealBoost algorithm and linguistic aggregation functions.	Instance space	Soft linguistic cardinality assumptions (e.g. some, many)
$\propto$SVM [20]	Solving a max-margin mixed-integer optimization problem, given predetermined instance label proportions, following alternating optimization or convex relaxation.	Instance space	Ratio-based assumption
MI-CRF [22]	Using a CRF where each node represents a bag which can take one of its instance as the value. In this model, all the bags are jointly classified based on unary instance classifiers and pairwise dissimilarity measurements	Instance space	Standard assumption
Structured Bag Models [21]	Using CRFs to model the bag structures and at the same time incorporating different MIL constraints. Learning is performed by minimizing an objective function with deterministic annealing approach	Instance space	Standard and Ratio-based assumptions
Generative Models for MIL [24]	Using Bayesian networks with different structures to learn generative models for MIL	Instance space	Standard assumption
Simple MI [25]	Mapping each bag to average of its instances and training a standard single-instance classifier.	Bag space	Metadata assumption
Histogram-Based Methods [7]	Finding a vocabulary of concepts by clustering the instances. Then, mapping each bag to a histogram vector of the concepts and finally train a single-instance classifier.	Bag space	Metadata assumption
DD-SVM [26] & MILES [14]	Mapping each bag to a vector built by the distances between the bag and instance prototypes of the DD algorithm. Next, classifying the vectors by the regular SVM (in DD-SV) or 1-norm SVM (in MILES).	Bag space	Metadata assumption
MI kernels [27]	Defining a number of MI kernels on bags and plug them into kernel methods.	Bag space	Metadata assumption
miGraph & MIGraph [30]	Mapping a bag into an undirected graph and designing a graph kernel. Next, classifying the bags by a kernel machine.	Bag space	Metadata assumption
Citation kNN [32]	Using a bag-to-bag distance in a modified nearest neighbor appraoch, where each bag is classified by majority voting among both citers and references.	Bag space	Nearest neighbor assumption (with B2B distance)
M-C2B [33]	Learning a robust and discriminative class-to-bag (C2B) distance for MIL by solving an $l_{2,1}$-norm minmax problem.	Bag space	Nearest neighbor assumption (with C2B distance)
Ours: Multi-Instance Markov Networks	Modeling bags using Markov networks with parameterized cardinality potentials so that different cardinality-based MI assumptions can be plugged into the models or even learned from data. Learning is formulated in a max-margin discriminative framework and solved with a non-convex cutting plane method.	Instance space + bag space	Any cardinality-based assumption + metadata assumption of bag-level features.

$\mathbb{R}^D$. Likewise, each bag might be globally described by another feature vector $\mathbf{X}$. For example, if the bag is an image, $\mathbf{X}$ can be a global bag-of-words feature vector extracted from the whole image. Another approach to construct $\mathbf{X}$ is using the prediction scores of other MIL methods[2] as a bag-level feature descriptor. Each instance $\mathcal{I}_i$ has also a hidden label y_i, and the collective binary instance labels of a bag are denoted by $\mathbf{y} = \{y_1, \cdots, y_m\}$. Given this notation, we propose a Markov network to define a scoring function over tuples $(\mathbf{X}, \mathbf{x} = \{\mathbf{x}_i\}_{i=1}^m, Y, \mathbf{y} = \{y_i\}_{i=1}^m)$. This function is used to predict the label of a test bag by inferring the bag and instance labels which maximize the scoring function, given the input feature vectors.

A graphical representation of the proposed Markov network is shown in Fig. 2. Each instance and its label are modeled by two nodes in a clique. The potential function of this clique specifies a classifier for an individual instance. A second clique contains all instance labels and the bag label. This clique is used to define what makes a bag positive or negative. Varying this clique potential will lead to different MI assumptions, and is the focus of our work. Finally, there is an optional clique potential between the global representation of the whole bag and the bag label.

2. In our experiments, we use MI-Kernel [27].

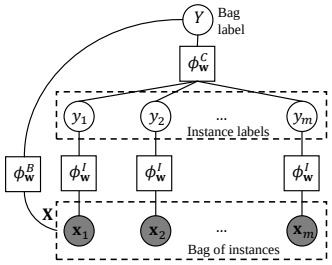


Fig. 2. Graphical illustration of the proposed model for binary multi-instance learning. Instance potential functions $\phi^I_{\mathbf{w}}(\mathbf{x}_i, y_i)$ relate instances $\mathbf{x}_i$ to labels y_i. A second clique potential $\phi^C_{\mathbf{w}}(\mathbf{y}, Y)$ relates all instance labels y_i to the bag label Y. There is also an optional potential function $\phi^B_{\mathbf{w}}(\mathbf{X}, Y)$, which relates the global representation of the bag to the bag label.

We define the scoring function on these cliques as:

$$f_{\mathbf{w}}(\mathbf{X}, \mathbf{x}, Y, \mathbf{y}) = \sum_i \phi^I_{\mathbf{w}}(\mathbf{x}_i, y_i) + \phi^C_{\mathbf{w}}(\mathbf{y}, Y) + \phi^B_{\mathbf{w}}(\mathbf{X}, Y), \tag{1}$$

where $\phi^I_{\mathbf{w}}(\mathbf{x}_i, y_i)$ represents the potential between each instance and its label, $\phi^C_{\mathbf{w}}(\mathbf{y}, Y)$ is the clique potential over all the instance labels and the bag label, and finally $\phi^B_{\mathbf{w}}(\mathbf{X}, Y)$ expresses the potential between the bag-level feature vector and the bag label. Note that the potential functions are parametrized by the learning weights $\mathbf{w}$. We explain the details of these potential functions as follows.

Instance-Label Potential $\phi^I_{\mathbf{w}}(\mathbf{x}_i, y_i)$: This potential function models the compatibility between the ith instance feature vector $\mathbf{x}_i$ and its label y_i. It is parametrized as:

$$\begin{aligned}\phi^I_{\mathbf{w}}(\mathbf{x}_i, y_i) &= \mathbf{w}_I^\top \, \mathbf{x}_i \, \mathbb{1}(y_i = 1) \\ &= \mathbf{w}_I^\top \, \Psi_I(\mathbf{x}_i, y_i).\end{aligned} \tag{2}$$

Labels Clique Potential $\phi^C_{\mathbf{w}}(\mathbf{y}, Y)$: This potential function models the relations between the instance labels and the bag label. Since the MIL problems are defined based on the number of positive and negative instances, we need to formulate this as a *cardinality* clique potential. Cardinality potentials are only a function of label counts – in this case, the counts of the positive and negative instances in the bag.

By modifying the form of the cardinality potential, we can encode different MI assumptions, which will be shown in the Section 3.2. Note that while for arbitrary clique potentials inference could be NP-complete, for cardinality potentials with binary variables exact and efficient inference algorithms exist. This leads to efficient algorithms for learning and prediction, which will be described in Section 4.

In order to define the cardinality potentials, we will use the notation m^+ and m^- for the counts of instance labels in $\mathbf{y}$ which are positive and negative, respectively. The complete clique potential depends on these counts, and the bag label Y. Thus, we describe this clique potential by parameterizing two different cardinality potential functions, one for positive bags ($C^+_{\mathbf{w}}$) and one for negative bags ($C^-_{\mathbf{w}}$).

$$\begin{aligned}\phi^C_{\mathbf{w}}(\mathbf{y}, Y) &= C_{\mathbf{w}}\left(m^+, m^-, Y\right) \\ &= C^+_{\mathbf{w}}\left(m^+, m^-\right) \mathbb{1}(Y = 1) \\ &\quad + C^-_{\mathbf{w}}\left(m^+, m^-\right) \mathbb{1}(Y = -1).\end{aligned} \tag{3}$$

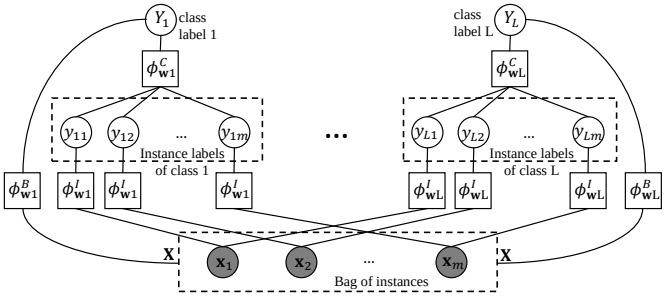


Fig. 3. Graphical illustration of the proposed model for multiclass multi-instance learning.

Bag-Label Potential $\phi^B_{\mathbf{w}}(\mathbf{X}, Y)$: This potential function gives a global model of a bag, which describes how the bag as a whole entity is classified. It is parametrized as:

$$\begin{aligned}\phi^B_{\mathbf{w}}(\mathbf{X}, Y) &= \mathbf{w}_B^\top \, \mathbf{X} \, \mathbb{1}(Y = 1) \\ &= \mathbf{w}_B^\top \, \Psi_B(\mathbf{X}, Y).\end{aligned} \tag{4}$$

3.1.2 Multiclass Classification

We can extend the binary model in Fig. 2 for multiclass classification. The proposed multiclass model is illustrated in Fig. 3. It can be observed that this network is formed by concatenation of the binary graphical model of each class. The main reason for this replication is that the inference of cardinality clique potentials is exact and efficient only for binary labels. To this end, first we represent the multiclass bag label $Y \in \{1, 2, \cdots, L\}$ by a binary vector $(Y_1, Y_2, \cdots, Y_L)$, where $Y_l = 1$ if $Y = l$ and $Y_l = -1$ if $Y \neq l$. In addition, for each class l, we have binary instance labels $\mathbf{y}_l = \{y_{l1}, \cdots, y_{lm}\}$ ($y_{li} \in \{+1, -1\}$, $i = 1, \cdots, m$), indicating which instances are from (or relevant to) the lth class and which instances are not. We also denote the collection of all instance labels of all classes by $\mathbf{y}$. Putting all this together, the scoring function of the tuple $(\mathbf{X}, \mathbf{x}, Y, \mathbf{y})$ for the proposed multiclass graphical model is defined by:

$$\begin{aligned}&f_{\mathbf{w}}(\mathbf{X}, \mathbf{x}, Y, \mathbf{y}) = \\ &\sum_{l=1}^{L} \left(\sum_i \phi^I_{\mathbf{w}l}(\mathbf{x}_i, y_{li}) + \phi^C_{\mathbf{w}l}(\mathbf{y}_l, Y_l) + \phi^B_{\mathbf{w}l}(\mathbf{X}, Y_l) \right),\end{aligned} \tag{5}$$

where, similar to the binary model, the instance-label potentials $\phi^I_{\mathbf{w}l}(\mathbf{x}_i, y_{li})$, the labels clique potential $\phi^C_{\mathbf{w}l}(\mathbf{y}_l, Y_l)$, and the bag-label potential $\phi^B_{\mathbf{w}l}(\mathbf{X}, Y_l)$ are defined as follows.

$$\begin{aligned}\phi^I_{\mathbf{w}l}(\mathbf{x}_i, y_{li}) &= \mathbf{w}_{Il}^\top \, \mathbf{x}_i \, \mathbb{1}(y_{li} = 1) \\ &= \mathbf{w}_{Il}^\top \, \Psi_I(\mathbf{x}_i, y_{li}).\end{aligned} \tag{6}$$

$$\begin{aligned}\phi^C_{\mathbf{w}l}(\mathbf{y}_l, Y_l) &= C_{\mathbf{w}l}\left(m_l^+, m_l^-, Y_l\right) \\ &= C^+_{\mathbf{w}l}\left(m_l^+, m_l^-\right) \mathbb{1}(Y_l = 1) \\ &\quad + C^-_{\mathbf{w}l}\left(m_l^+, m_l^-\right) \mathbb{1}(Y_l = -1).\end{aligned} \tag{7}$$

$$\begin{aligned}\phi^B_{\mathbf{w}}(\mathbf{X}, Y_l) &= \mathbf{w}_{Bl}^\top \, \mathbf{X} \, \mathbb{1}(Y_l = 1) \\ &= \mathbf{w}_{Bl}^\top \, \Psi_B(\mathbf{X}, Y_l).\end{aligned} \tag{8}$$

The following section defines functions $C^+_{\mathbf{w}l}$ and $C^-_{\mathbf{w}l}$ that lead to a variety of MIL models.

3.2 The Proposed Models of Multi-Instance Classification

In this section, we use our proposed Markov network to model MIL with different MI assumptions.

3.2.1 Multiple Instance Markov Network (MIMN)

This network models the multi-class multi-instance classification with the standard MI assumption, i.e., a bag of class label l has at least one instance from the lth class. Thus, in this model, the labels clique potential for each possible class label $l \in \{1, \cdots, L\}$ is given by

$$C^+_{\mathbf{w}l}(0, m) = -\infty \tag{9}$$
$$C^+_{\mathbf{w}l}(m^+_l, m - m^+_l) = w^+_{Cl} \quad m^+_l = 1, \cdots, m \tag{10}$$
$$C^-_{\mathbf{w}l}(0, m) = w^-_{Cl} \tag{11}$$
$$C^-_{\mathbf{w}l}(m^+_l, m - m^+_l) = -\infty \quad m^+_l = 1, \cdots, m. \tag{12}$$

This clique potential states that in a bag of class label l it is impossible to have no instance from the lth class (9), and there is the same potential of having one or more than one instance from the target class (10). However, if the bag label is not equal to l, none of the instances should be from this class (11) & (12). One could set w^+_{Cl} and w^-_{Cl} to a constant value (e.g. 0)[3], but we generally treat them as the model parameters and show how to learn them in Section 4.2.

3.2.2 Ratio-constrained Multiple Instance Markov Network (RMIMN)

Ratio-constrained MIL extends the notion of labeled bags in MIL based on instance labels proportions. In RMIMN, each bag of class label l contains at least a certain portion of instances from class l. For example, at least 30% of the instances should be from the lth class in a bag with label l. To encode this MI assumption with our proposed Markov network, we only need to refine the functions $C^+_{\mathbf{w}l}$ and $C^-_{\mathbf{w}l}$:

$$\begin{aligned}
&C^+_{\mathbf{w}l}(m^+_l, m - m^+_l) = -\infty \quad 0 \le \frac{m^+_l}{m} < \rho \\
&C^+_{\mathbf{w}l}(m^+_l, m - m^+_l) = w^+_{cl} \quad \rho \le \frac{m^+_l}{m} \le 1 \\
&C^-_{\mathbf{w}l}(m^+_l, m - m^+_l) = w^-_{cl} \quad 0 \le \frac{m^+_l}{m} < \rho \\
&C^-_{\mathbf{w}l}(m^+_l, m - m^+_l) = -\infty \quad \rho \le \frac{m^+_l}{m} \le 1,
\end{aligned} \tag{13}$$

where ρ indicates the threshold proportion of relevant instances in a bag. The interesting case is $\rho = 0.5$, where we can learn models with majority voting assumption.

3.2.3 Generalized Multiple Instance Markov Network (GMIMN)

GMIMN allows a very flexible notion of labeled bags. We allow the proportion of relevant and irrelevant instances in bags to be a learned parameter, discovered from the data. The MIL model will learn which fractions of instances tend to be of the target class in a bag of that class. This network provides a very general model for multiple instance learning and is parametrized by:

$$\begin{aligned}
&C^+_{\mathbf{w}l}(0, m) = -\infty \\
&C^+_{\mathbf{w}l}(m^+_l, m - m^+_l) = \sum_{k=1}^{K} w^+_{kl} \mathbb{1}\left(\frac{k-1}{K} < \frac{m^+_l}{m} \le \frac{k}{K}\right) \\
&\qquad\qquad m^+_l = 1, \cdots, m \\
&C^-_{\mathbf{w}l}(m^+_l, m - m^+_l) = \sum_{k=1}^{K} w^-_{kl} \mathbb{1}\left(\frac{k-1}{K} \le \frac{m^+_l}{m} < \frac{k}{K}\right) \\
&\qquad\qquad m^+_l = 0, \cdots, m-1 \\
&C^-_{\mathbf{w}l}(m, 0) = -\infty.
\end{aligned} \tag{14}$$

where K determines the number of weighted segments of a bag. This model divides the bag size into K equal parts, and the weight of each segment w_{kl} determines how important it is that the number of relevant instances (i.e., the instances from class l) be placed inside that interval. In other words, these learning weights specify the importance or impact of different witness ratios for labeling a bag. Large values of K provide more detailed models of bag definition by learning cardinality-based measures with finer resolution, while low values of K define a coarser model of bag. So, by controlling the granularity, this parameter is set in a trade-off between training accuracy and generalization ability[4].

The constraints $C^+_{\mathbf{w}l}(0, m) = -\infty$ and $C^-_{\mathbf{w}l}(m, 0) = -\infty$ are the only required prior information in this model, which break the symmetry between positive and negative bags and enforce at least one instance of a positive bag is positive and one instance of a negative bag is negative. Note that since this model is very general and unconstrained, it is vulnerable to overfitting (especially for multi-class classification) and requires careful training practices[5] to achieve successful results.

3.2.4 Linearity of the Models

In Section 3.1, we showed that the *instance-label* potentials and the *bag-label* potential are linear functions of the learning weights $\mathbf{w}$ (See equations (6) and (8)). Here, we demonstrate the linearity of the cardinality-based *labels clique* potential with respect to $\mathbf{w}$. Consequently, the whole model score would be a linear function of the learning parameters.

Given $C^+_{\mathbf{w}l}$ and $C^-_{\mathbf{w}l}$ defined for any of the MIMN, RMIMN, or GMIMN models, the labels clique potential for each class label (e.g., the lth class) can be written as:

$$\phi^C_{\mathbf{w}l}(\mathbf{y}_l, Y_l) = \mathbf{w}^\top_{Cl} \Psi_C(\mathbf{y}_l, Y_l) + g_C(\mathbf{y}_l, Y_l), \tag{15}$$

where $\mathbf{w}_{Cl}$ represents the concatenation of the learning parameters in $C^+_{\mathbf{w}l}$ and $C^-_{\mathbf{w}l}$, while $\Psi_C(\mathbf{y}_l, Y_l)$ and $g_C(\mathbf{y}_l, Y_l)$ are functions independent of $\mathbf{w}l$, which are specified by aggregation of the indicator functions.

Now, by integrating all the potential functions of the multi-class Markov network, the scoring function introduced in (5) is reduced to the following linear expression:

$$f_{\mathbf{w}}(\mathbf{X}, \mathbf{x}, Y, \mathbf{y}) = \mathbf{w}^\top \Psi(\mathbf{X}, \mathbf{x}, Y, \mathbf{y}) + \sum_l g_C(\mathbf{y}_l, Y_l), \tag{16}$$

3. Our experimental explorations show that setting these parameters to zero usually leads to satisfactory results

4. In the experiments of this paper, we cross-validate on the values $K = 3$, $K = 5$, and $K = 10$ to roughly estimate this parameter.

5. Examples of good practices are smart initialization of the learning weights (e.g. using the weights learned by MIMN model) and early stopping on the training iterations by monitoring the validation error.

where

$$\Psi(\mathbf{X},\mathbf{x},Y,\mathbf{y}) = [\sum_i \Psi_I(\mathbf{x}_i,y_{1i})^\top,\cdots,\sum_i \Psi_I(\mathbf{x}_i,y_{Li})^\top, \Psi_C(\mathbf{y}_1,Y_1)^\top,\cdots,\Psi_C(\mathbf{y}_L,Y_L)^\top, \Psi_B(\mathbf{X},Y_1)^\top,\cdots,\Psi_B(\mathbf{X},Y_L)^\top]^\top. \quad (17)$$

This linearity property facilitates parameter learning with gradient-based methods, which will be explained in Section 4.2.

4 Inference and Learning

The MIL models above define scoring functions $f_{\mathbf{w}}$ which consider counts of instance labels in a bag (see Eq. (5)). Using this, we can define a scoring function for assigning the bag label Y to a bag with bag feature $\mathbf{X}$ and instance features $\mathbf{x}$ by maximum a posteriori (MAP) inference of the Markov network over the hidden instance labels:

$$F_{\mathbf{w}}(\mathbf{X},\mathbf{x},Y) = \max_{\mathbf{y}} f_{\mathbf{w}}(\mathbf{X},\mathbf{x},Y,\mathbf{y}). \quad (18)$$

Below, we describe how to efficiently solve this inference problem for the cardinality-based cliques we defined above. Using this inference technique, learning can be performed using a max-margin criterion, as in the Latent SVM approach [16].

Classification of a new test bag can be done in a similar manner. We can predict the bag label by simply running inference, enumerating all possible Y and taking the maximum scoring bag label:

$$Y^\star = \arg\max_Y F_{\mathbf{w}}(\mathbf{X},\mathbf{x},Y). \quad (19)$$

4.1 Inference

The inference problem is to find the best set of instance labels of all class labels $\mathbf{y}^\star = \{\mathbf{y}_1^\star,\mathbf{y}_2^\star,\cdots,\mathbf{y}_L^\star\}$ given the input feature vectors for the data $\{\mathbf{X},\mathbf{x}\}$ and the bag label Y. Using (5) and (7), the inference problem in (18) can be written as

$$\mathbf{y}^\star = \max_{\mathbf{y}} \sum_{l=1}^{L} (\sum_i \phi^I_{\mathbf{w}l}(\mathbf{x}_i,y_{li}) + C_{\mathbf{w}l}(m_l^+,m_l^-,Y_l)). \quad (20)$$

However, the instance labels of each class are conditionally independent from instance labels of other classes, given the input feature vectors and the bag label fixed. Thus, the original inference problem of all instance labels is decomposed and reduced to inference of the instance labels of each class, separately:

$$\mathbf{y}_l^\star = \max_{\mathbf{y}_l} \sum_i \phi^I_{\mathbf{w}l}(\mathbf{x}_i,y_{li}) + C_{\mathbf{w}l}(m_l^+,m_l^-,Y_l). \quad (21)$$

This problem is the standard problem of inferring a probabilistic graphical model with cardinality clique potentials [38]. This class of PGMs is specified by two parts: the sum of individual node potentials and a clique potential over all the nodes which only depends on the counts of the nodes which get specific labels. In our models, we only work with binary node labels (i.e., $y_{li} \in \{+1,-1\}$), for which there exists an exact inference algorithm with $O(m\log m)$ time complexity[6]. The inference algorithm is as follows. First, sort the instances in decreasing order of $\phi^I_{\mathbf{w}l}(\mathbf{x}_i,+1) - \phi^I_{\mathbf{w}l}(\mathbf{x}_i,-1)$. Then, for $k=0,\cdots,m$, compute s_k^l, the sum of the top-k instance potentials $\phi^I_{\mathbf{w}l}(\mathbf{x}_i,+1) - \phi^I_{\mathbf{w}l}(\mathbf{x}_i,-1)$ plus the clique potential $C_{\mathbf{w}l}(k,m-k,Y_l)$. Finally, find $k_l^\star$ which gets the largest s_k^l, and inference is accomplished by assigning the top $k_l^\star$ instances to positive labels and the rest to negative labels. Repeating this algorithm for each class label, the full inference of (20) takes $O(L\,m\log m)$ time.

4.2 Learning

Let the training set be given by $\{(\mathbf{X}^1,\mathbf{x}^1,Y^1),\cdots,(\mathbf{X}^N,\mathbf{x}^N,Y^N)\}$, and the goal is to train the Markov models by learning the parameters $\mathbf{w}$. Inspired by the relation to latent SVM, we formulate the learning problem as minimizing the regularized hinge loss function:

$$\begin{aligned} &\min_{\mathbf{w}} \sum_{n=1}^{N} (\mathcal{L}^n - \mathcal{R}^n) + \frac{\lambda}{2}\|\mathbf{w}\|^2 \\ &\text{where } \mathcal{L}^n = \max_Y \max_{\mathbf{y}} (\Delta(Y,Y^n) + f_{\mathbf{w}}(\mathbf{X}^n,\mathbf{x}^n,Y,\mathbf{y})), \\ &\mathcal{R}^n = \max_{\mathbf{y}} f_{\mathbf{w}}(\mathbf{X}^n,\mathbf{x}^n,Y^n,\mathbf{y}), \\ &\Delta(Y,Y^n) = \begin{cases} 1 & \text{if } Y \neq Y^n \\ 0 & \text{if } Y = Y^n. \end{cases} \end{aligned} \quad (22)$$

One approach to solve this problem approximately is the iterative algorithm of alternating between inference of the latent variables and optimization of the model parameters. So, the first step estimates the instance labels and the second step learns a standard SVM classifier given the estimated instance labels. It can be shown using this approach for the binary MIMN model leads to the mi-SVM algorithm [11].

However, we use the non-convex regularized bundle method (NRBM) [41] to directly solve the optimization problem in (22). It has been shown that NRBM has a fast convergence rate compared to the state-of-the-art non-convex optimiztion methods [42]. This method iteratively makes an increasingly accurate piecewise quadratic approximation of the objective function. At each iteration, a new linear cutting plane is obtained via the subgradient of the objective function and added to the piecewise quadratic approximation. To use this algorithm, the principal issue is to compute the subgradients $\partial_{\mathbf{w}}\mathcal{L}^n(\mathbf{w})$ and $\partial_{\mathbf{w}}\mathcal{R}^n(\mathbf{w})$. For this purpose, we need to know the subgradient of the network scoring function, i.e., $\partial_{\mathbf{w}} f_{\mathbf{w}}(\mathbf{X},\mathbf{x},Y,\mathbf{y})$.

Following the linear expression derived in (16), it is simple to show that

$$\partial_{\mathbf{w}} f_{\mathbf{w}}(\mathbf{X},\mathbf{x},Y,\mathbf{y}) = \Psi(\mathbf{X},\mathbf{x},Y,\mathbf{y}), \quad (23)$$

Using equations (22) and (23), it can be shown that $\partial_{\mathbf{w}}\mathcal{L}^n(\mathbf{w}) = \Psi(\mathbf{X}^n,\mathbf{x}^n,Y^\star,\mathbf{y}^\star)$, where $(\mathbf{y}^\star,Y^\star)$ is the solution to the inference problem:

$$\max_Y \max_{\mathbf{y}} (\Delta(Y,Y^n) + f_{\mathbf{w}}(\mathbf{X}^n,\mathbf{x}^n,Y,\mathbf{y})). \quad (24)$$

6. For non-binary node labels, there exist only approximate inference algorithms. See [38] for more details.

This inference problem can be solved using the algorithm in Section 4.1. In summary, we enumerate all possible Y, and for each fixed Y we find $\mathbf{y}$ by doing inference on the resulting graphical model (which has cardinality clique potentials). Then, the Y with the highest value gives the predicted bag label $Y^\star$.

In the same way, it can be shown that $\partial_{\mathbf{w}}\mathcal{R}^n(\mathbf{w}) = \Psi(\mathbf{X}^n, \mathbf{x}^n, Y^n, \mathbf{y}^\star)$, where $\mathbf{y}^\star$ is the solution to the inference problem:

$$\max_{\mathbf{y}} f_{\mathbf{w}}(\mathbf{X}^n, \mathbf{x}^n, Y^n, \mathbf{y}). \tag{25}$$

5 Experiments

In this section, we show the performance of the proposed framework in different classification tasks. First, the MIMN model is evaluated on binary and multiclass MIL benchmark datasets. Next, the extended models are applied to the two challenging computer vision tasks of cyclist helmet recognition and human activity recognition to show that the flexibility in the portion of positives in a bag can lead to improved classification accuracy.

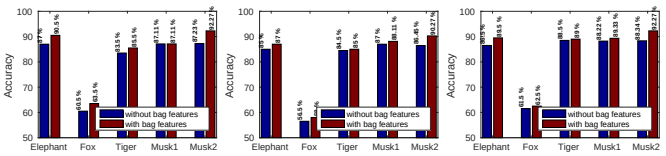
5.1 Benchmark Datasets

In this section, we evaluate our proposed MIMN model on MIL benchmark datasets to demonstrate it can achieve the state-of-the-art performance on standard datasets.

5.1.1 Binary Benchmarks

We evaluate the MIMN model on five popular binary MIL datasets[7]. These benchmark datasets are the *Elephant*, *Fox*, *Tiger* image data sets [11] and *Musk1* and *Musk2* drug activity prediction data sets [8]. In the image data sets, each bag represents an image, and the instances inside the bag represent 230-D feature vectors of different segmented blobs of the image. These datasets contain 100 positive and 100 negative bags. In the Musk datasets, each bag describes a molecule, and the instances inside the bag represent 166-D feature vectors of the low-energy configurations of the molecule. Musk1 has 47 positive bags and 45 negative bags with about 5 instances per bag. Musk2 has 39 positive bags and 63 negative bags with variable number of instances in a bag, ranging from 1 to 1044 (average 64 instances per bag).

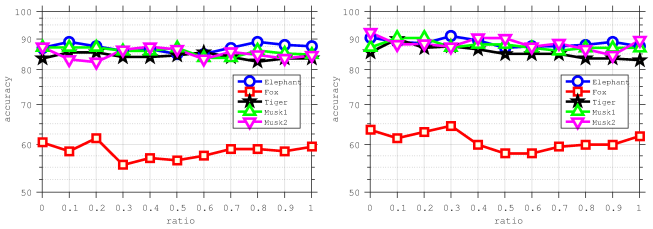
In all experiments of this section, the instance features have been extended by approximate explicit *intersection kernel* mapping [43], and the bag features have been constructed by the prediction scores of the MI-Kernel method [27] with RBF kernel. In addition, the features have been preprocessed by scaling the original features to the range $[0, 1]$. At each experimental trial, we run the non-convex cutting plane algorithm with all the learning weights initialized to 0 (except bag features[8]) and at most 100 iterations. The regularization parameter λ was roughly optimized on the 10-fold cross-validation accuracy by grid search in a set of predetermined values (1, 10, and 100). The averaged classification accuracies for the MIMN model on different datasets are shown in Fig. 4. It can be observed that combining the MIMN model with the bag features helps to improve the results.



(a) MIMN (b) RMIMN ($\rho = 0.5$) (c) GMIMN ($K = 5$)

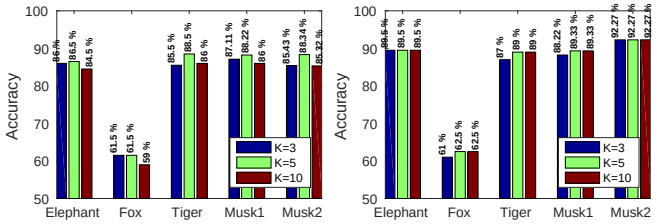
Fig. 4. Evaluating the classification performance of the proposed models on binary benchmark datasets.

We also illustrate the performance of RMIMN model with different values of ρ in Fig. 5. This figure shows that RMIMN is somehow robust to the value of ρ on these datases. The reason might be because there is no inherent ratio-constrained assumption in these benchmark datasets. We show the merit of RMIMN in Section 5.2 and 5.3 when the experiment are performed on real computer vision tasks with intuitive ratio-constrained assumptions. Next, we demonstrate GMIMN results with different values of K in Fig. 6. It is shown that this value influences the performance of the GMIMN model, and it is beneficial to set the proper value by doing cross-validation. However, note that when the bag features are integrated, the model becomes more robust to K.



(a) without bag features (b) with bag features

Fig. 5. Classification accuracy on binary benchmark datasets using RMIMN with different value of ρ.



(a) without bag features (b) with bag features

Fig. 6. Classification accuracy on binary benchmark datasets using RMIMN with different value of ρ.

Finally, we compare the MIMN models with the state-of-the-art MIL methods in Table 2. The performance of the methods varies depending on the data set. However, our proposed models are always among the best methods. More specifically, MIMN and GMIMN achieve the best accuracy on the Elephant, Fox, Tiger, and Musk2 data sets, compared to the other methods.

7. The original data sets are available online at http://www.cs.columbia.edu/~andrews/mil/datasets.html.

8. Since the bag features are the MI-Kernel prediction scores, we initialize the corresponding weights to small positive values, e.g. 0.1, so that the first iteration of the algorithm will be the same as MI-Kernel.

TABLE 2
Comparison between state-of-the-art MIL methods on the binary MIL benchmark datasets. The best and second best results are highlighted in bold and italic face respectively.

Method	Elephant	Fox	Tiger	Musk1	Musk2
MIMN	*89*	**64**	*86*	87	**92**
RMIMN ($\rho = 0.5$)	87	59	85	88	**92**
GMIMN ($K = 5$)	**90**	63	**89**	89	**92**
mi-SVM [11]	82	58	79	87	84
MI-SVM [11]	81	59	84	78	84
MI-Kernel [27]	84	60	84	88	*89*
γ-rule SVM [44]	84	63	81	88	85
SetMaxRBM^{XOR} [23]	88	60	83	84	84
MIRealBoost [19]	83	63	73	**91**	77
MIForest [15]	84	**64**	82	85	82
SVR-SVM [45]	85	63	80	88	85
MIGraph [30]	85	61	82	*90*	90
miGraph [30]	87	62	*86*	*90*	90
MILES [14]	81	62	80	88	83
AW-SVM [17]	82	**64**	83	86	84
AL-SVM [17]	79	63	78	86	83
EM-DD [10]	78	56	72	85	85

TABLE 3
Comparison between state-of-the-art MIL methods on the COREL image datasets. The numbers show the average accuracy over 5 trials and the corresponding 95% intervals.

Method	1000-Image	2000-Image
MIMN	**85.6** $\pm$ 0.5	71.6 $\pm$ 1.0
RMIMN ($\rho = 0.5$)	85.2 $\pm$ 0.6	**72.1** $\pm$ 0.6
GMIMN ($K = 10$)	84.9 $\pm$ 0.4	70.9 $\pm$ 0.7
MI-Kernel [27]	84.1 $\pm$ 0.6	69.1 $\pm$ 0.7
MKSVM-MIL [46]	85.2 $\pm$ 1.1	71.3 $\pm$ 1.2
MILES [14]	81.5 $\pm$ 3.0	68.7 $\pm$ 1.4
DD-SVM [26]	74.7 $\pm$ 1.6	67.5 $\pm$ 0.8
MissSVM [47]	78.0 $\pm$ 2.2	65.2 $\pm$ 3.1
MI-SVM [11]	74.7 $\pm$ 1.6	54.6 $\pm$ 1.5

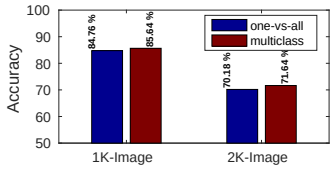


Fig. 7. Comparison between classification accuracy of the proposed multiclass MIMN and binary MIMN with one-vs-all technique.

5.1.2 Multiclass Benchmarks

In this section, we evaluate the multiclass extension of the MIMN model for image categorization on the COREL dataset. We work on the 1000-image and 2000-image datasets[9] [14], which contain ten and twenty categories with 100 images per category. Each image is represented as a bag of instances, where the instances are the ROIs (Region of Interests) described by nine features (representing color, shape, and energy).

The experiments are performed with the same setup as in Section 5.1.1, i.e. extending and scaling the instance features and making MI-Kernel bag features. Also, the same experimental routine as described in [14] was used: the images of each category are split into half for training and test, and the experiment on each dataset is repeated five times. The results are provided in Table 3 and compared with other MIL methods. Note that the accuracy of MI-Kernel is based on our implementation, and for the other methods the numbers are reported from [46]. As seen in the table, MIMN models are competitive with the state-of-the-art methods [10].

To show the contribution of our proposed multiclass formulation, Fig. 7 compares multiclass MIMN with binary MIMN wrapped by exhaustive one-vs-all technique. Our empirical evaluations show that the muliclass model obtain higher classification accuracy[11].

5.2 Cyclist Helmet Recognition

In this section, we use our proposed models to address a binary video classification task. This problem is illustrated in Fig. 8. Given an automatically-obtained cyclist trajectory, we must determine whether the cyclist is wearing a helmet or not. One can treat this as a MIL problem – each frame is an instance, and the trajectory forms a bag. The bag (trajectory) should be classified as containing a helmet-wearing cyclist or not. However, the standard MI assumption or traditional supervised learning approaches (e.g. classify each instance and majority vote) cannot easily handle this problem. Because of imperfection in tracking, it is unlikely that all the instances in a positive bag are truly positive – some will not be well centered on the cyclist's head due to jitter, regardless of the tracker used. Traditional supervised learning would have many corrupted positive instances of helmet-wearing cyclists. Standard MI assumption would not make full use of the training data, since each track would very likely have more than one positive instance.

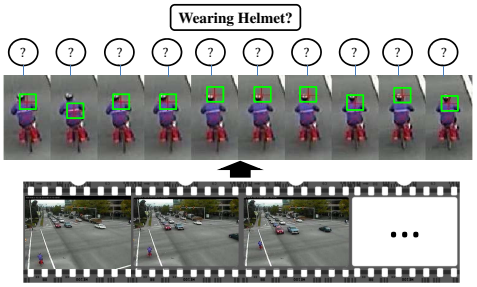


Fig. 8. Cyclist helmet classification – is she wearing helmet? how many positives are in this bag? An automatic cyclist detector/tracker is run, with head position estimate in green rectangle. Data instances are features defined on the head position estimates, bags aggregate these over a track.

5.2.1 Experimental Setup

We work with cyclist trajectories automatically extracted from video data. The data are collected for a busy 4-legged intersection with vehicles, pedestrians, and cyclists, over a two-day period. Kanade-Lucas-Tomasi feature tracking

9. The original data sets are available online at http://www.miproblems.org/datasets/corel.

10. In all our experimental studies we found that GMIMN is not very successful for multiclass classification. The reason tend to be the loose and weak assumption in GMIMN as well as the large number of free parameters in the multiclass version, which makes the model overfit to the training data.

11. Also, considering the same number of iterations, multiclass MIMN is faster in practice.

TABLE 4
Results of the experiments on cyclist helmet classification problem.

Method	Accuracy %
SVM-AtLeastOne	58.33
SVM-Majority	79.17
mi-SVM	62.50
MIMN	58.33
RMIMN ($\rho = 0.5$)	91.67
GMIMN ($K = 5$)	87.50

and trajectory clustering are used to extract moving objects. These clusters are then automatically classified (vehicle, pedestrian, cyclist) by analyzing speed profiles (e.g. the pedalling cadence).

We chose a dataset of 24 cyclist tracks for our experiments – 12 wearing helmets and 12 not. The head location is estimated using background subtraction upon the tracks. We describe each frame of a track using texton histograms [48] in a region of size 20×20 around the head position (chosen after empirically examining other features). We report the results of helmet classification using leave-one-out cross-validation on this dataset.

We use the proposed models in Section 3 to classify the cyclist tracks. We also compare this approach with non-MIL methods. In the non-MIL approach, all frames from positive and negative training videos are put together and labelled according to their video labels. Next, a standard SVM classifier [49] is trained and used to predict each frame label of the test videos. Finally, the bag label is predicted by one of the following criteria:

- SVM-AtLeastOne: The bag label is positive if at least one of the instance labels is positive.
- SVM-Majority: The bag label is specified by the majority voting of the instance labels.

5.2.2 Experimental Results

For our proposed algorithms, we run the non-convex cutting plane algorithm with all the learning weights initialized to 0 and at most 100 iterations. For all the algorithms the regularization parameter was estimated by grid search on the cross-validation accuracy. The average classification accuracy of each method is shown in Table 4. We include mi-SVM as an additional baseline. The results of the RMIMN model with different ρ values are demonstrated in Fig. 9.

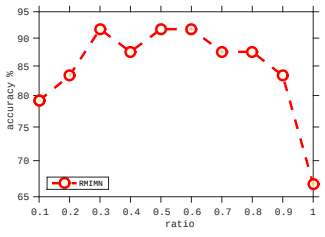


Fig. 9. Cyclist helmet recognition accuracy with RMIMN model and different values of the parameter ρ.

It can be observed that the classification accuracy of SVM-AtLeastOne, MIMN, and mi-SVM are quite low. This shows that the traditional classification approach (used in SVM-AtLeastOne) and the standard MI assumption (used in MIMN and mi-SVM) are very ineffective in this problem. The standard MI assumption fails because it is very likely that at least one of the instances in a negative bag is classified as positive, and consequently most of the negative bags are assigned positive labels. This problem is due to the imperfection in the classifier and low-quality visual representation of the cyclist's head in the video. However, it is clearly evident that SVM-Majority, RMIMN (with most ρ values), and GMIMN are more robust to these defects. The results show that RMIMN (with $\rho = 0.5$) outperforms all the other methods. Also, it is shown that GMIMN has competitive performance. It learns the multi-instance relation properly without any prior knowledge of the ambiguity level (e.g., parameter ρ) and classifies the videos successfully.

5.3 Group Activity Recognition

In this section, we show the application of the proposed cardinality-based multi-instance models for group activity recognition. We run experiments on two datasets: nursing home dataset [2] and collective activity dataset [1].

5.3.1 Nursing Home Dataset

In this section, our method is evaluated for activity recognition in a nursing home. The dataset we use [2] provides scenes in which the individuals might be performing different actions such as walking, standing, sitting, bending, or falling. However, the goal is to detect the "fall" event, i.e., if any person is falling or not in a scene. Thus, we use the proposed binary MIMN model to encode that *at least one* of the individuals is falling in a positive scene. Fig. 10 illustrates the problem of fall scene detection in the nursing home dataset.

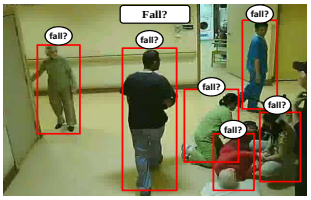


Fig. 10. An example of "Fall" scene from the nursing home dataset. We model this problem as a multi-instance learning problem, where each individual is represented as an instance.

The dataset has 22 video clips (12 clips for training and 8 clips for test) with 2990 annotated frames, where about one third of them are assigned the "fall" activity label. We use the same features and experimental settings as used in [2]. The results in terms of classification accuracy are shown in Table 5. We compare our method with global bag-of-words method and the spatial structured models in [2]. Note that because of the significant class size imbalance, mean per-class accuracy is a more valid performance criterion. It can be observed that our proposed MIMN model outperforms the others. It is an intuitive outcome because of the problem definition (at least one person is falling in a fall scene). The results of the RMIMN model with different values of ρ, shown in Fig. 11, also follow this intuition.

TABLE 5
Comparison of different methods on the nursing home dataset in terms of classification accuracy (CA) and mean per-class accuracy (MPCA). We used the same features and experimental settings as in [2].

Method	CA	MPCA
Global bag-of-words with SVM [2]	52.6	53.9
Latent SVM with unconnected graph [2]	58.6	56.0
Latent SVM with tree-structured graph [2]	64.1	60.6
Latent SVM with complete graph [2]	70.0	63.1
Latent SVM with optimized graph structure [2]	71.2	65.0
MIMN (ours)	76.1	**66.2**
RMIMN ($\rho = 0.5$)	75.3	60.6
GMIMN ($K = 10$)	**77.1**	65.5

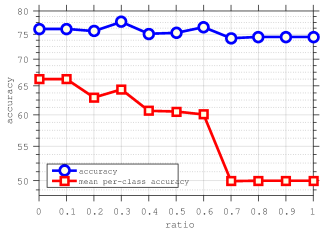


Fig. 11. Recognition accuracy on nursing home dataset with RMIMN model and different values of the parameter ρ.

5.3.2 Collective Activity Dataset

In this section, we study the application of the proposed models in the multiclasss classification task of collective activity recognition. The collective activity dataset [1] comprises 44 videos (about 2500 video frames) of *crossing*, *waiting*, *queuing*, *walking*, and *talking*. The goal is to classify the collective activity in each video frame, where the collective activity commonly tends to be the action that the majority of people in the scene are doing. For this purpose, each frame scene is modeled as a bag of people described by the *action context* feature descriptors[12] proposed in [2]. The MIL representation of this problem is shown in Fig. 12. In our experiments, the same experimental setup is followed as explained in [2], i.e., the same 1/3 of the video clips were selected for test and the rest for training. We use our proposed RMIMN model with $\rho = 0.5$ to encode *majority* multi-instance assumption on the action labels. The results are shown in Table 6 and compared with the following methods: (1) SVM with global bag-of-words model and (2) spatial latent structured models in [2].

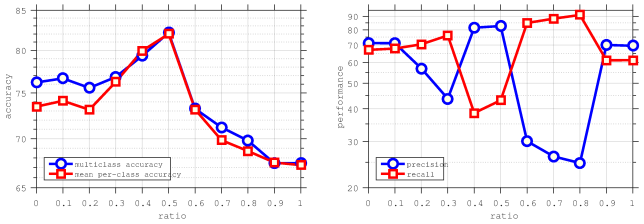
TABLE 6
Comparison of different methods on collective activity dataset in terms of multi-class classification accuracy (MCA) and mean per-class accuracy (MPCA). We used the same settings as in [2].

Method	MCA	MPCA
Global bag-of-words with SVM [2]	70.9	68.6
Latent SVM with optimized graph [2]	79.7	78.4
MIMN	76.2	73.5
RMIMN ($\rho = 0.5$)	**82.2**	**82.0**
GMIMN ($K = 3$)	72.3	70.0

12. Note that this feature descriptor is built on a spatio-temporal context region around any individual. So it encodes the spatio-temporal information in the action and its context. By using our multi-instance model, the spatio-temporal and cardinality information are combined.

(a) (b)

Fig. 12. Two examples from the collective human activity recognition dataset. (a) shows a scene where the collective activity is waiting while (b) shows a similar scene but the collective activity is crossing. The intuition is that the collective activity tends to be the action that majority of people are doing. We model this problem as a MIL problem, where the goal is to recognize the collective activity in the scene by inferring the hidden action each person is doing. We use our proposed RMIMN model to encode the majority multi-instance assumption.



(a) Bag-level classification (b) Instance-level prediction

Fig. 13. Performance of RMIMN with different value of ρ on collective activity dataset.

Our proposed RMIMN Model can achieve the best results, even compared to the structure-optimized spatial model in [2], by incorporating the cardinality relations into the problem. More specifically, Fig. 13 illustrates the results of the RMIMN model for both bag-level classification and instance-level prediction with different values of ρ. It can be observed that as expected, the highest bag classification accuracy is obtained with $\rho = 0.5$ (Fig. 13(a)).

Instance-level prediction results (in terms of precision and recall averaged over all action classes) are shown in Fig. 13(b). This analysis lends further insight, though note that instance-level predictions are dependent on bag-level classification. At high bag-level accuracy ($\rho = 0.4$ and $\rho = 0.5$), recall is around 40% and 50% with high precision, which is expected: the model predicts enough instance-level positives to satisfy the bag being positive.

Outside this range, at lower bag-level accuracies, the picture is different. As ρ increases, more instances are predicted as positive instances and the instance-level recall is likely to increase. However, the chance of misspredition (especially in false positive bags) also increases and the precision decreases. However, at some point ($\rho = 0.9$), where there is a stringent constraint for classifying a bag as positive, the number of false negative bags tends to increase. Consequently, the number of positive instances shrinks, and the precision is enhanced.

Finally, visualizations of some example recognition results are provided in Fig. 14.

6 Conclusion

We proposed a novel graphical framework for both binary and multiclass multi-instance learning based on Markov

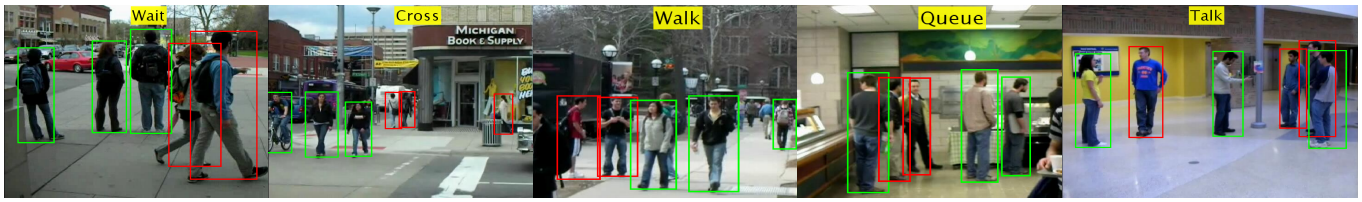


Fig. 14. Visualization of some recognition results of the proposed method. Each figure is annotated by the predicted collective activity. Also each individual is represented by a colored bounding box. If an individual is involved in the predicted collective activity, the bounding box is green, otherwise red (In fact, these colors are used to illustrate predicted instance labels – green for positive label and red for negative label). For example, in the first figure from left, three people are waiting and two people are walking (passing by in the street). In the second figure, three people are crossing and the others are walking. In the third figure, all the people are walking except two people who are talking. Note that the instance labels are not always correctly predicted. For example in the fourth figure although all the people are involved in the queuing activity, two of them are incorrectly labeled by red. Also, in the last figure three people are incorrectly labeled. It seems that because of our weakly supervised learning framework (where we only incorporate the whole scene collective activity label in the max-margin learning formulation and model the individual action labels with hidden variables), the resulting model is sometimes conservative in predicting the instance labels and tries to detect just enough positive instances to predict the whole scene collective activity correctly.

networks and latent max-margin discriminative training. This framework is flexible and can model any cardinality-based multi-instance assumptions. Thus, it is more robust to the amount of labeling ambiguity (i.e. true positive instances) in the bags. Specifically, it can be helpful in vision applications which exhibit imperfect annotation or ambiguous feature representations. Further, it can be used to model visual recognition problems with intrinsic cardinality relations (e.g. group activity recognition).

The experiments showed that learning and encoding the degree of ambiguity in the classifier can influence the accuracy of classification. We used the proposed framework for binary classification of cyclists with and without helmet. We also evaluated the performance of the multiclass models on the collective activity recognition problem. These are challenging problems, where the traditional supervised learning and standard MI assumptions fail. However, the extended ratio-based models enhance classification performance by encoding more general and robust MI assumptions and mining the degree of ambiguity.

The proposed graphical framework is flexible and can be easily extended or modified. For example, it can be modified for multi-label multi-instance learning, where a bag can take more than one label. Also, the model can be extended by defining new potential functions. For example, pairwise potential functions could be defined over neighbouring instance labels to model spatial or temporal relations between the instances. Finally, this framework can be adapted for individual classification from group statistics with applications in privacy-preserving data mining [50, 51, 20].

REFERENCES

[1] W. Choi, K. Shahid, and S. Savarese, "What are they doing?: Collective activity classification using spatio-temporal relationship among people," in *9th International Workshop on Visual Surveillance*, 2009.

[2] T. Lan, Y. Wang, W. Yang, S. N. Robinovitch, and G. Mori, "Discriminative latent models for recognizing contextual group activities," *IEEE T-PAMI*, vol. 34, no. 8, pp. 1549–1562, 2012.

[3] W. Choi and S. Savarese, "A unified framework for multi-target tracking and collective activity recognition," in *ECCV*, 2012, pp. 215–230.

[4] M. R. Amer, D. Xie, M. Zhao, S. Todorovic, and S. C. Zhu, "Cost-sensitive top-down/bottom-up inference for multiscale activity recognition," in *ECCV*, 2012.

[5] H. Hajimirsadeghi, J. Li, G. Mori, M. Zaki, and T. Sayed, "Multiple instance learning by discriminative training of markov networks," in *UAI*, 2013.

[6] J. Foulds and E. Frank, "A review of multi-instance learning assumptions," *Knowledge Engineering Review*, vol. 25, no. 1, p. 1, 2010.

[7] J. Amores, "Multiple instance classification: Review, taxonomy and comparative study," *Artificial Intelligence*, vol. 201, pp. 81–105, 2013.

[8] T. Dietterich, R. Lathrop, and T. Lozano-Pérez, "Solving the multiple instance problem with axis-parallel rectangles," *Artificial Intelligence*, vol. 89, no. 1-2, pp. 31–71, 1997.

[9] O. Maron and T. Lozano-Pérez, "A framework for multiple-instance learning." MORGAN KAUFMANN PUBLISHERS, 1998, pp. 570–576.

[10] Q. Zhang and S. A. Goldman, "Em-dd: An improved multiple-instance learning technique," in *NIPS*, 2001, pp. 1073–1080.

[11] S. Andrews, I. Tsochantaridis, and T. Hofmann, "Support vector machines for multiple-instance learning," in *NIPS*, 2002.

[12] O. L. Mangasarian and E. W. Wild, "Multiple instance classification via successive linear programming," *Journal of Optimization Theory and Applications*, vol. 137, no. 3, pp. 555–568, 2008.

[13] R. Bunescu and R. Mooney, "Multiple instance learning for sparse positive bags," in *ICML*, 2007, pp. 105–112.

[14] Y. Chen, J. Bi, and J. Wang, "Miles: Multiple-instance learning via embedded instance selection," *IEEE T-PAMI*, vol. 28, no. 12, pp. 1931–1947, 2006.

[15] C. Leistner, A. Saffari, and H. Bischof, "Miforests: Multiple-instance learning with randomized trees," in *ECCV*. Springer, 2010, pp. 29–42.

[16] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part based models," *IEE T-PAMI*, vol. 32, no. 9, pp. 1627–1645, 2010.

[17] P. Gehler and O. Chapelle, "Deterministic annealing for multiple-instance learning," in *AISTATS*, 2007.

[18] W. Li, L. Duan, D. Xu, and I. Tsang, "Text-based image retrieval using progressive multi-instance learning," in *ICCV*, 2011, pp. 2049–2055.
[19] H. Hajimirsadeghi and G. Mori, "Multiple instance real boosting with aggregation functions," in *ICPR*, 2012.
[20] F. X. Yu, D. Liu, S. Kumar, T. Jebara, and S.-F. Chang, "$\propto$ svm for learning with label proportions," in *ICML*, 2013.
[21] J. Warrell and P. H. Torr, "Multiple-instance learning with structured bag models," in *Energy Minimization Methods in Computer Vision and Pattern Recognition*. Springer, 2011, pp. 369–384.
[22] T. Deselaers and V. Ferrari, "A conditional random field for multiple-instance learning," in *ICML*, 2010.
[23] J. Louradour and H. Larochelle, "Classification of sets using restricted boltzmann machines," in *UAI*, 2011.
[24] T. Adel, R. Urner, B. Smith, D. Stashuk, and D. J. Lizotte, "Generative multiple-instance learning models for quantitative electromyography," in *UAI*, 2013, pp. 1–11.
[25] L. Dong, "A comparison of multi-instance learning algorithms," Ph.D. dissertation, The University of Waikato, 2006.
[26] Y. Chen and J. Z. Wang, "Image categorization by learning and reasoning with regions," *The Journal of Machine Learning Research*, vol. 5, pp. 913–939, 2004.
[27] T. Gärtner, P. Flach, A. Kowalczyk, and A. Smola, "Multi-instance kernels," in *ICML*, 2002, pp. 179–186.
[28] T. Gärtner, "Kernel-based feature space transformation in inductive logic programming," *Master's thesis, University of Bristol*, 2000.
[29] J. T. Kwok and P.-M. Cheung, "Marginalized multi-instance kernels." in *IJCAI*, 2007, pp. 901–906.
[30] Z. Zhou, Y. Sun, and Y. Li, "Multi-instance learning by treating instances as non-iid samples," in *ICML*, 2009, pp. 1249–1256.
[31] P. Viola, J. Platt, and C. Zhang, "Multiple instance boosting for object detection," in *NIPS*, 2006, pp. 1417–1424.
[32] J. Wang and J.-D. Zucker, "Solving multiple-instance problem: A lazy learning approach," in *ICML*, 2000, pp. 1119–1125.
[33] H. Wang, F. Nie, and H. Huang, "Robust and discriminative distance for multi-instance learning," in *CVPR*. IEEE, 2012, pp. 2919–2924.
[34] H. Wang, H. Huang, F. Kamangar, F. Nie, and C. H. Ding, "Maximum margin multi-instance learning," in *NIPS*, 2011, pp. 1–9.
[35] M. Guillaumin, J. Verbeek, and C. Schmid, "Multiple instance metric learning from automatically labeled bags of faces," in *ECCV*. Springer, 2010, pp. 634–647.
[36] H. Wang, F. Nie, and H. Huang, "Learning instance specific distance for multi-instance classification." in *AAAI*, 2011, pp. 507–512.
[37] F. Nie, H. Huang, X. Cai, and C. H. Ding, "Efficient and robust feature selection via joint 2, 1-norms minimization," in *NIPS*, 2010, pp. 1813–1821.
[38] R. Gupta, A. Diwan, and S. Sarawagi, "Efficient inference with cardinality-based clique potentials," in *ICML*. ACM, 2007, pp. 329–336.
[39] D. Tarlow, K. Swersky, R. Zemel, R. Adams, and B. Frey, "Fast exact inference for recursive cardinality models," in *Uncertainty in Artificial Intelligence (UAI-12)*, 2012.
[40] L. Duan, W. Li, I. Tsang, and D. Xu, "Improving web image search by bag-based re-ranking," *IEEE Trans. Image Process.*, vol. 20, no. 11, pp. 3280–3290, 2011.
[41] T. Do and T. Artières, "Large margin training for hidden markov models with partially observed states," in *ICML*. ACM, 2009, pp. 265–272.
[42] T.-M.-T. Do and T. Artières, "Regularized bundle methods for convex and non-convex risks," *JMLR*, vol. 13, no. 1, pp. 3539–3583, 2012.
[43] A. Vedaldi and A. Zisserman, "Efficient additive kernels via explicit feature maps," *IEEE T-PAMI*, vol. 34, no. 3, pp. 480–492, 2012.
[44] Y. Li, D. M. Tax, R. P. Duin, and M. Loog, "Multiple-instance learning as a classifier combining problem," *Pattern Recognition*, vol. 46, no. 3, pp. 865–874, 2013.
[45] F. Li and C. Sminchisescu, "Convex multiple-instance learning by estimating likelihood ratio," *NIPS*, pp. 1360–1368, 2010.
[46] D. Li, J. Wang, X. Zhao, Y. Liu, and D. Wang, "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.
[47] Z.-H. Zhou and J.-M. Xu, "On the relation between multi-instance learning and semi-supervised learning," in *ICML*, 2007, pp. 1167–1174.
[48] J. Malik, S. Belongie, T. K. Leung, and J. Shi, "Contour and texture analysis for image segmentation," *IJCV*, vol. 43, no. 1, pp. 7–27, 2001.
[49] C. Chang and C. Lin, "Libsvm: a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 2, no. 3, p. 27, 2011.
[50] N. Quadrianto, A. Smola, T. Caetano, and Q. Le, "Estimating labels from label proportions," *The Journal of Machine Learning Research*, vol. 10, pp. 2349–2374, 2009.
[51] S. Rueping, "Svm classifier estimation from group probabilities," in *ICML*, 2010, pp. 911–918.

Hossein Hajimirsadeghi received the Ph.D. degree in Computer Science from Simon Fraser University, Canada. He received his M.Sc. and B.Sc. in Electrical Engineering from University of Tehran, Iran in 2010 and 2008, respectively. He is currently a researcher in Oracle Labs, Canada. His research interests are in machine learning and computer vision, including probabilistic graphical models, kernel learning, deep learning, and action/activity recognition.

Greg Mori received the Ph.D. degree in Computer Science from the University of California, Berkeley in 2004. He received an Hon. B.Sc. in Computer Science and Mathematics with High Distinction from the University of Toronto in 1999. He is currently a professor in the School of Computing Science at Simon Fraser University. Dr. Mori's research interests are in computer vision, and include object recognition, human activity recognition, human body pose estimation.