Multi-View Multi-Instance Learning Based on Joint Sparse Representation and Multi-View Dictionary Learning

Bing Li, Chunfeng Yuan, Weihua Xiong, Weiming Hu, Houwen Peng, Xinmiao Ding, Steve Maybank

Abstract—In multi-instance learning (MIL), the relations among instances in a bag convey important contextual information in many applications. Previous studies on MIL either ignore such relations or simply model them with a fixed graph structure so that the overall performance inevitably degrades in complex environments. To address this problem, this paper proposes a novel multi-view multi-instance learning algorithm (M^2IL) that combines multiple context structures in a bag into a unified framework. The novel aspects are: (i) we propose a sparse ε-graph model that can generate different graphs with different parameters to represent various context relations in a bag, (ii) we propose a multi-view joint sparse representation that integrates these graphs into a unified framework for bag classification, and (iii) we propose a multi-view dictionary learning algorithm to obtain a multi-view graph dictionary that considers cues from all views simultaneously to improve the discrimination of the M^2IL. Experiments and analyses in many practical applications prove the effectiveness of the M^2IL.

Index Terms—multi-instance learning, multi-view, sparse representation, dictionary learning

1 Introduction

As a variant of supervised learning, multi-instance learning (MIL) represents a sample by a bag of several instances instead of a single one. It only gives each bag, not each instance, a discrete or real-valued label. Starting from the original work of Dietterich et al [1], MIL has been used in many applications [1] [2] [3].

1.1 Related Work

Recent decades have witnessed great progress in MIL algorithms [5] [6] [7]. We roughly divide existing MIL methods into two categories, "independent MIL methods (IMIL)" and "contextual MIL methods (CMIL)". These two categories differ in the way that the relations among instances in a bag are treated.

The IMIL methods treat all the instances from a bag as independently and identically distributed (i.i.d.). These methods can be further divided into generative IMIL and discriminative IMIL. Axis-Parallel Rectangles (APR) [1], Diverse Density (DD) [8], Expectation-Maximization (EM) version of Diverse Density (EM-DD) [9], Generalized EM-based Diverse Density (GEM-DD) [10] are all in the generative IMIL category. MIL problems can also be tackled in a discriminative manner by adapting standard supervised learning approaches. The methods of this type learn a classifier that separates positive and negative bags. The work falling in this category can be traced back to the citation k-nearest neighbor (CKNN) method [11]. Wang et al [12] propose a maximum margin MIL algorithm based on a type of class-to-bag distance. The support vector machine (SVM) is also introduced to solve MIL, resulting in a plethora of SVM-based MIL algorithms, including multi-instance kernels (MI-kernel) [13], support vector machine for MIL (MI-SVM and mi-SVM) [14], DD-SVM [3], MIL via embedded instance selection (MILES) [2], MIL with instance selection (MILIS) [7], Fast Bundle Algorithm for MIL [15] and others [16] [17].

Fig. 1. An example of airplane recognition using MIL: (A) the background of "sky" provides an important context for the recognition of "airplane", (B) the background of "mountain" is not a useful cue for the recognition of "airplane"

The CMIL methods differ from the IMIL ones in that they treat the instances in a bag as non-i.i.d by taking into account the interplay of the instances. Zhou and Xu [17] point out that the relations among instances in a bag convey important structural information in many applications. They propose two CMIL methods, MIGraph and miGraph [18], which define the relations among instances in a bag with a ε-graph [19] and apply SVM with two graph kernel functions to bag classification. Since then, CMIL has attracted many researchers attention. Song et al. [20] apply the miGraph to identify user attributes in social network services. Li et al [21] propose a CMIL algorithm by adding a contextual constraint on a Fuzzy SVM. Zhang et al [23] present a CMIL framework for structured data classification.

- *B. Li, C. Yuan, W. Xiong, H. Peng, and X. Ding are with the National Laboratory of Pattern Recognition (NLPR), Institute of Automation, CAS, Beijing, 100190, China. E-mail: {bli, cfyuan}@nlpr.ia.ac.cn*
- *W. Hu is with CAS Center for Excellence in Brain Science and Intelligence Technology, National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences; University of Chinese Academy of Sciences. E-mail: wmhu@nlpr.ia.ac.cn*
- *S. Maybank is with Department of Computer Science and Information Systems, Birkbeck College, UK.*

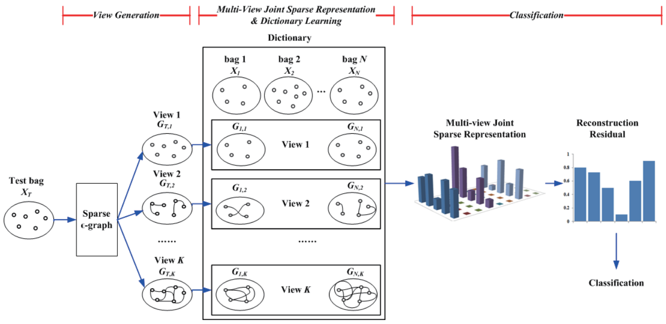

Fig. 2. The framework of the M^{2}IL.

Although the existing MIL methods from both categories are claimed to achieve good performance in many tasks, they have two limitations:

(i) The relations among instances are defined to be either independent or contextual for all bags. However, in many applications, we cannot simply pre-define the instances in a bag as independent or not. Taking Figure 1 as an example, an image is viewed as a bag of objects ("instances") and we would like to recognize the concept of "airplane" using MIL. In Figure 1(A), the background of "sky" can provide an important contextual cue and should not be neglected. The background of "mountain" in Figure 1(B) has no contextual relationship with "airplane".

(ii) It is still a difficult problem to define the relations among instances in complex and varied environments. In most existing CMIL methods, the instances relations are described by a ε-graph with a fixed ε value. It is unreasonable for these methods to represent diverse contextual relations using only one type of graph.

1.2 Our Work

To circumvent the limitations of the existing MIL methods and inspired by the idea of multi-view learning [24], we propose a multi-view multi-instance learning algorithm (M^{2}IL) based on a joint sparse representation. The contributions of this paper are summarized as follows:

(i) It proposes a sparse ε-graph model that integrates ε-graph and ℓ_1-graph models into a unified framework, and can generate different graphs in a systematic way with different parameters.

(ii) It proposes a multi-view multi-instance learning model (M^{2}IL) based on a joint sparse representation and graph structures. The "multi-view here is defined as a series of inherent contextual structures among instances in a bag. These structures are represented by undirected graphs generated via the proposed sparse ε-graph model, and are integrated into a unified multi-view joint sparse representation framework for bag classification.

(iii) It proposes a novel multi-view dictionary learning algorithm for the M^{2}IL. Different from the existing dictionary learning algorithms [32] [33], the proposed algorithm learns a multi-view graph dictionary by considering cues from all views simultaneously.

2 Overview of M^{2}IL

Before giving an overview of the proposed M^{2}IL, we briefly review the formal definition of the MIL. Let χ denote the instance space. We are given a data set $\{(X_1, y_1), ..., (X_i, y_i), ..., (X_N, y_N)\}$, where $X_i = \{x_{i,1}, x_{i,2}, ..., x_{i,n_i}\} \subseteq \chi$ is called a bag and $y_i \in \eta = \{+1, -1\}$ is the label of the bag X_i. Here $x_{i,j} \in R^p$ (suppose that each $x_{i,j}$ is normalized to have unit ℓ^2 norm) is called an instance in bag X_i. If there exists $m \in \{1, ..., n_i\}$ such that $x_{i,m}$ is a positive instance, then X_i is a positive bag and $y_i = 1$; otherwise X_i is a negative bag and $y_i = -1$. The value of m is always unknown. That is, for any positive bag, we only know that there is at least one positive instance in it, but cannot figure out which ones they are from. The goal of MIL is to learn a classifier to predict the labels of unseen bags.

In order to consider the relations among instances in a bag, this paper proposes a novel multi-view multi-instance learning (M^{2}IL), in which a series of graphs are added to each bag to represent the contextual relations among the instances. Figure 2 illustrates the basic idea of the M^{2}IL. It contains three key steps: view generation, multi-view joint sparse representation and dictionary learning, and classification.

View Generation. In M^{2}IL, for any bag X_i, we first construct a set of K undirected graphs $\Gamma_i = \{G_{i,1}, G_{i,2}, ..., G_{i,K}\}$ where each $G_{i,k}$ is defined by $G_{i,k} =< X_i, \mathbf{M}_{i,k} >$ with all the instances in X_i as the vertices and an edge set represented by an adjacency matrix $\mathbf{M}_{i,k} \in R^{n_i \times n_i}$. If there is an edge between $x_{i,a}$ and $x_{i,b}$, then $\mathbf{M}_{i,k}(a, b) = \mathbf{M}_{i,k}(b, a) = 1$, otherwise $\mathbf{M}_{i,k}(a, b) = \mathbf{M}_{i,k}(b, a) = 0$. All the K graphs are generated by the proposed sparse ε-graph model with K different choices of parameter values. The graphs can be viewed as different contextual structures among instances in the bag X_i.

Multi-View Joint Sparse Representation and Dictionary Learning. Given a graph set Γ_i, the traditional MIL is extended to the M^{2}IL by explicitly including Γ_i in the training data as $\{(X_1, \Gamma_1 =< G_{1,1}, ..., G_{1,K} >, y_1), ..., (X_i, \Gamma_i =< G_{i,1}, ..., G_{i,K} >, y_i), ..., (X_N, \Gamma_N =< G_{N,1}, ..., G_{N,K} >, y_N)\}$. The labels in the M^{2}IL can be binary or multiple, as $y_i \in \{1, 2, ..., C\}$ for C classes.

To solve the M^{2}IL problem, this paper proposes a multi-view joint sparse representation framework. It is essentially a sparse classifier aiming at reconstructing the k^{th} graph of a bag with the graphs from the k^{th} view of a learned dictionary. It has been observed that an effective dictionary usually leads to a more compact representation and better performance in many applications [27] [32] [33]. Therefore, we design a multi-view dictionary learning algorithm based on graph kernels to learn a discriminative dictionary for each class from training bags.

Classification. For a test bag with a set of graphs Γ_T and an unknown label y_T as $(X_T, \Gamma_T =< G_{T,1}, ..., G_{T,K} >, y_T)$, each of the K graphs is reconstructed using the learned multi-view dictionary under the M^{2}IL framework. The reconstruction residual from all the K views is used to predict the label y_T.

3 Sparse ε-graph for View Generation

In the view generation step, the undirected graphs $\Gamma_i = \{G_{i,1}, G_{i,2}, ..., G_{i,K}\}$ are generated for each bag X_i. Zhou et al [18] use the ε-graph to model the local manifold structure among instances in a bag for MIL. The ε-graph is defined with the pair-wise Euclidean distance and a global threshold, making it sensitive to noise. Cheng et al [25] construct a ℓ_1-graph in which the edge between any two vertices is determined by a sparse representation. However, locality must lead to sparsity but not necessary vice versa [26]. We propose a novel sparse ε-graph model to avoid the disadvantages of ℓ_1-graph and ε-graph models. The proposed sparse ε-graph model can specify the edges locally and adaptively by adding a distance constraint to the ℓ_1-graph. It is a unified framework that can generate different kinds of graphs with different parameters.

3.1 Sparse ε-graph

This section discusses how to construct the graph $G_{i,k} =< X_i, \mathbf{M}_{i,k} >$ for each bag based on the sparse ε-graph model. Without loss of generality and for simplicity, we remove the index k in $G_{i,k} =< X_i, \mathbf{M}_{i,k} >$ and write it as $G_i =< X_i, \mathbf{M}_i >$.

Before detailing the sparse ε-graph, we briefly show how to define the structure of a bag X_i using the ℓ_1-graph [25]. Given a bag X_i, the ℓ_1-graph constructs the graph $G_i =< X_i, \mathbf{M}_i >$ based on the sparse representation [27]. Considering an vertex $x_{i,j}$ and its edges to the other vertices $\mathbf{U} = [u_1, u_2, ..., u_{n_i-1}] = [x_{i,1}, x_{i,2}, ...x_{i,j-1}, x_{i,j+1}, ..., x_{i,n_i}] \in R^{p\times(n_i-1)}$, the ℓ_1-graph is to find a sparse vector of coefficients $\alpha \in R^{n_i-1}$, such that $x_{i,j} \approx \mathbf{U}\alpha = \sum_{k=1}^{n_i-1} u_k\alpha_k$. The vector is obtained by solving the following sparse representation objective function,

$$\min_{\alpha} \|x_{i,j} - \mathbf{U}\alpha\|^2 + \lambda\|\alpha\|_1, \tag{1}$$

where the first term of (1) is the linear reconstruction error, and the second term controls the sparsity of α through a regularization coefficient λ. Larger values of λ imply sparser values of α. The edges from $x_{i,j}$ to other instances are determined by values of α. If the coefficient $\alpha_k \neq 0$, then the element $\mathbf{M}_i(j,k) = 1$; otherwise, $\mathbf{M}_i(j,k) = 0$.

It is not guaranteed that the neighbors (that is $\alpha_k \neq 0$) to $x_{i,j}$ in the ℓ_1-graph are also near to $x_{i,j}$ in the Euclidean distance [26]. To circumvent this limitation, we add a Euclidean distance constraint to (1). We first define a weight matrix $\mathbf{D}$ based on the Euclidean distances from $x_{i,j}$ to the other vertices as:

$$\begin{aligned}\mathbf{D} = diag(&\varpi(\|x_{i,j} - x_{i,1}\|), ..., \varpi(\|x_{i,j} - x_{i,j-1}\|),\\ &\varpi(\|x_{i,j} - x_{i,j+1}\|), ..., \varpi(\|x_{i,j} - x_{i,n_i}\|)),\end{aligned} \tag{2}$$

where $\varpi(\|x_{i,j} - x_{i,k}\|) > 0$ is a monotone increasing function of the Euclidean distance $\|x_{i,j} - x_{i,k}\|$. Then we add the weight matrix $\mathbf{D}$ into (1) as:

$$\min_{\alpha} \|x_{i,j} - \mathbf{U}\alpha\|^2 + \lambda\|\mathbf{D}\alpha\|_1, \tag{3}$$

where $\lambda\|\mathbf{D}\alpha\|_1$ is the regularization item that considers both sparsity of α and the Euclidean distances from $x_{i,j}$ to the other vertices. The goal of (3) is to find those vertices with lower distance values to $x_{i,j}$ to reconstruct it. Although the function $\varpi(\|x_{i,j} - x_{i,k}\|)$ can be defined as any monotone increasing function, we define it as a piecewise constant one to simplify the optimization of (3), as:

$$\varpi(\|x_{i,j} - x_{i,k}\|) = \begin{cases} 1, & \|x_{i,j} - x_{i,k}\| \leq \varepsilon \\ \infty, & \|x_{i,j} - x_{i,k}\| > \varepsilon \end{cases}, \tag{4}$$

where ε is a threshold controlling the value of the corresponding element in $\mathbf{D}$(1 or ∞). According to (3) and (4), if an instance $x_{i,k}$ has $\|x_{i,j} - x_{i,k}\| > \varepsilon$, the weight for $x_{i,k}$ in matrix $\mathbf{D}$ is ∞, the instance $x_{i,k}$ will not be selected to reconstruct the instance $x_{i,j}$, and the corresponding coefficient value α_k will be 0. In other words, there is no edge linking $x_{i,j}$ and $x_{i,k}$ in the graph $G_i =< X_i, \mathbf{M}_i >$.

With the definition in (4), (3) can be simply solved by selecting those elements in $\mathbf{U}$ having distances less than ε from $x_{i,j}$ for sparsely representing $x_{i,j}$. It includes 3 major steps: (i) Set $\alpha_k = 0, \text{if}(\ \|u_k - x_{i,j}\|\ > \ \varepsilon)$. (ii) The remaining elements ($\{u_k |\, \|u_k - x_{i,j}\| \leq \varepsilon\}$) are used to compose an instance matrix $\mathbf{U}'$, and then used to sparsely represent $x_{i,j}$ ($\min_{\beta} \|x_{i,j} - \mathbf{U}'\beta\|^2 + \lambda\|\beta\|_1$) based on (1). The coefficient vector β can be obtained using existing sparse representation algorithms. (iii) Finally, the value of α_k (where $\|u_k - x_{i,j}\| \leq \varepsilon$) is set as the corresponding value in β. The detailed implement of the sparse ε-graph is given in Algorithm 1. More analysis about the spare ε-graph can be found in Appendix A.

Algorithm 1 sparse ε-graph construction.

Input: A bag in MIL as $X_i = \{x_{i,1}, x_{i,2}, ..., x_{i,n_i}\}$, parameter $\theta =< \lambda, \varepsilon >$.
Initialize: The matrix $\mathbf{M}_i$ for bag X_i as $\mathbf{M}_i = \mathbf{0}$.
for $j = 1 \to n_i$, $t = 1 \to n_i$ **do**
  Set $\mathbf{U} = [X_i \backslash x_{i,j}]$.
  Solve (3), obtain the value of sparse code α.
  If $t < j$, set $\mathbf{M}_i(j,t) = |\alpha_t|$;
  If $t = j$, set $\mathbf{M}_i(j,t) = 1$;
  If $t > j$, set $\mathbf{M}_i(j,t) = |\alpha_{t-1}|$;
end for
Set $\mathbf{M}_i = (\mathbf{M}_i + \mathbf{M}_i^T)/2$.
for $j = 1 \to n_i$, $t = 1 \to n_i$ **do**
  if $\mathbf{M}_i(j,t) \neq 0$ **then**
    set $\mathbf{M}_i(j,t) = 1$.
  end if
end for
Output: an undirected graph $G =< X_i, \mathbf{M}_i >$.

3.2 View Generation Using Sparse ε-graph

Using the proposed sparse ε-graph, we can generate different graphs with various parameters $< \lambda, \varepsilon >$, as:

(i) $\varepsilon = 0$, Independent Set. In this situation, all the elements in $\mathbf{D}$ are ∞, and the solution for α is $\alpha = \mathbf{0}$. The generated graph is a set of independent vertices without edges.

(ii) $\varepsilon \geq 2$, ℓ_1-graph. Since $\|x_{i,j} - x_{i,k}\| \leq 2$, all the diagonal elements in $\mathbf{D}$ are 1. Now (3) is equivalent to (1), and the sparse ε-graph reduces to the ℓ_1-graph.

(iii) $0 < \varepsilon < 2$, $\lambda \to 0$, ε-graph. When λ is very small, the sparsity constraint in (3) is weak, and the coefficient vector α becomes dense. The resulting graph approximates to a ε-graph.

(iv) $0 < \varepsilon < 2$, $\lambda > 0$, sparse ε-graph. In this situation, the sparsity constraint in (3) is emphasized, resulting in a smaller number of vertices selected in reconstruction of $x_{i,j}$.

For each bag X_i, we can generate K different graphs $\Gamma_i = \{G_{i,1}, G_{i,2}, ..., G_{i,K}\}$ using different parameter settings $\{<\lambda_1,\varepsilon_1>,<\lambda_2,\varepsilon_2>,...,<\lambda_K,\varepsilon_K>\}$ to represent the inner contextual structures of X_i from different views.

4 Multi-View Joint Sparse Representation and Dictionary Learning

4.1 Multi-View Joint Sparse Representation

The sparse representation-based classification (SRC) has been successfully used in many applications [27] [28]. We extend the SRC to a multi-view joint sparse representation-based classification model for MIL. After obtaining the K graphs for each bag, given any bag with K graphs and its label $(X_\tau, \Gamma_\tau =< G_{\tau,1},...,G_{\tau,K} >, y_\tau)$, the multi-view joint sparse representation is to represent the k^{th} graph of the bag sparsely, using the k^{th} graphs of dictionaries. Since the graph structure cannot be directly used for sparse representation, we apply a feature mapping function $\varphi : G \mapsto R^d$ to map a graph G to a high dimensional feature space as: $G \mapsto \varphi(G)$ and define the sparse representation in the mapped feature space. The feature vectors obtained from the k^{th} graphs of all the training bags are arranged as the columns of a feature matrix $\mathbf{V}^k = [\varphi(G_{1,k}), \varphi(G_{2,k}), ..., \varphi(G_{N,k})] \in R^{d\times N}$. For convenience of description, we sort the graphs in $\mathbf{V}^k$ according to the corresponding bag labels, as $\mathbf{V}^k = [\mathbf{V}_1^k, \mathbf{V}_2^k, ..., \mathbf{V}_C^k]$, where $\mathbf{V}_j^k = [\varphi(G_{1,k}), \varphi(G_{2,k}), ..., \varphi(G_{N_j,k})] \in R^{d\times N_j}$ denotes the graphs of all the training bags in the j^{th} class, and N_j is the number of training bags in the j^{th} class ($N_1 + N_2 + ... + N_C = N$). Similar to SRC, we let $\mathcal{D}^k \in R^{d\times M}$ be a sought dictionary with M atoms for the k^{th} view and let $\mathcal{D} = \{\mathcal{D}^1, \mathcal{D}^2, ..., \mathcal{D}^K\}$ be the set of all the dictionaries for all the views that can be learned from all training samples $\mathbf{V}^k, (k = 1,...,K)$. Each dictionary $\mathcal{D}^k \in R^{d\times M}$ is composed of all the class-specific sub-dictionaries as $\mathcal{D}^k = [\mathcal{D}_1^k, \mathcal{D}_2^k, ..., \mathcal{D}_C^k]$ where $\mathcal{D}_j^k \in R^{d\times M_j}$ is the sub-dictionary of the j^{th} class with M_j atoms and $M_1 + M_2 + ... + M_C = M$. The sparse representation of the bag $(X_\tau, \Gamma_\tau =< G_{\tau,1},...,G_{\tau,K} >, y_\tau)$ view can be written as

$$\min_{\mathbf{W}^k} \left\|\varphi(G_{\tau,k}) - \mathcal{D}^k\mathbf{W}^k\right\|_2^2 + \gamma\left\|\mathbf{W}^k\right\|_1, \tag{5}$$

where $\mathbf{W}^k \in R^M$ is the sparse representation coefficient vector for $\varphi(G_{\tau,k})$ and γ is a regularization coefficient. Given the coefficient vector $\mathbf{W}^k$, (5) expresses how to sparsely reconstruct each of the K graphs of the bag X_τ. If we consider the sparse representations from all K views, the sparse representation can be written as

$$\min_{\mathbf{W}} \sum_{k=1}^{K} \left(\left\|\varphi(G_{\tau,k}) - \mathcal{D}^k\mathbf{W}^k\right\|_2^2 + \gamma\left\|\mathbf{W}^k\right\|_1\right), \tag{6}$$

where $\mathbf{W} = [\mathbf{W}^1, ..., \mathbf{W}^K] \in R^{M\times K}$ is the matrix obtained by stacking the K columns of coefficient vectors $\{\mathbf{W}^k\}$. Each row of the matrix $\mathbf{W}$ is the coefficient vector associated with a training bag over K views, while each column of the matrix $\mathbf{W}$ is the coefficient vector associated with all the M atoms in a dictionary over a view.

From the viewpoint of multi-task learning, the ℓ_1-norm regularization in (6) is essentially defined on K independent sparse representations. It has two obvious drawbacks: (i) It does not take into account the relationships among the graph structures from different views. As a result, the solution does not benefit from any combination of multiple views. (ii) It uses all the atoms in the dictionaries independently and neglects the labels of them during the reconstruction procedure.

Yuan and Yan [29] show that reconstruction based on independent views and independent dictionary atoms is unreliable and sensitive to noise in many practical situations. They further point out that the reconstruction can benefit from prior knowledge about the relationships among dictionaries. To combine the strength of multiple views, we replace the ℓ_1-norm regularization with a joint one by imposing a class-level sparsity-inducing ℓ_2-norm regularization [29]. The intuition of this extension is that the introduced regularization can jointly select a few common classes to represent a bag over graph structures from multiple views in the task of bag classification. To this end, let $\mathbf{W}_j \in R^{M_j\times K}$ denote a sub-matrix of the coefficient matrix $\mathbf{W}$ corresponding to the dictionary $\mathcal{D}_j^k \in R^{d\times M_j}$ in the j^{th} class. We now have $\mathbf{W} = [(\mathbf{W}_1)^T, (\mathbf{W}_2)^T, ..., (\mathbf{W}_C)^T]^T \in R^{(M_1+M_2+...+M_C)\times K}$. To combine the strength of all the dictionaries within the j^{th} class over all views, we first apply the ℓ_2-norm over $\mathbf{W}_j$ (i.e. $\|\mathbf{W}_j\|_F$), and then apply the ℓ_1-norm across the ℓ_2-norm of the $\mathbf{W}_j$ (i.e. $\sum_{j=1}^{C}\|\mathbf{W}_j\|_F$) to promote sparsity to allow a small number of classes to be involved during the joint sparse representation. Thus, we arrive at the following class-level group joint sparse representation as

$$\min_{\mathbf{W}} \left\{\frac{1}{2}\sum_{k=1}^{K}\left\|\varphi(G_{\tau,k}) - \mathcal{D}^k\mathbf{W}^k\right\|_2^2 + \gamma\sum_{j=1}^{C}\|\mathbf{W}_j\|_F\right\}. \tag{7}$$

The class-specific multi-view joint sparse representation in (7) simultaneously considers both multiple views and class prior in reconstructing the bag X_τ.

4.2 Multi-View Dictionary Learning

According to the objective function in (7), we should first learn the dictionary $\mathcal{D} = \{\mathcal{D}^1, \mathcal{D}^2, ..., \mathcal{D}^K\}$ from training data. Inspired by the success of the meta-face algorithm for face recognition [31] that learns a face dictionary for each class separately, we also learn the class-specific sub-dictionary $\mathcal{D}_j = \{\mathcal{D}_j^1, \mathcal{D}_j^2, ..., \mathcal{D}_j^K\}, (j = 1,...,C)$ for each class, separately.

The most important property of the class-specific sub-dictionary $\mathcal{D}_j = \{\mathcal{D}_j^1, \mathcal{D}_j^2, ..., \mathcal{D}_j^K\}$ is self-expressiveness inner class [30] [31], meaning that the class-specific sub-dictionary provides a basis pool that can well sparsely represent all the training samples in the j^{th} class over K views. Let $\theta_j = \{X_i | y_i = j\}$ denote all the training bags in the j^{th} class, and let $\mathbf{P}_i \in R^{M_j\times K}$ be the reconstruction coefficient matrix of the i^{th} training bag in the j^{th} class based on the dictionary $\mathcal{D}_j$. The objective function of the class-specific dictionary learning can be defined as [30] [31]:

$$\arg\min_{\mathcal{P},\mathcal{D}_j} \sum_{i=1,X_i\in\theta_j}^{N_j} \left\{\frac{1}{2}\sum_{k=1}^{K}\left\|\varphi(G_{i,k}) - \mathcal{D}_j^k[\mathbf{P}_i]^k\right\|_2^2 + \gamma\|\mathbf{P}_i\|_{2,1}\right\}, \tag{8}$$

where $\mathcal{P} = \{\mathbf{P}_1, \mathbf{P}_2, ..., \mathbf{P}_{N_j}\}$ is the list of coefficient matrices of all the training bags in the j^{th} class, $[\mathbf{P}_i]^k$ is the

k^{th} column of $\mathbf{P}_i$ indicating the coefficient vector associated with k^{th} view, and $\|\bullet\|_{2,1}$ is the $\ell_{2,1}$-norm that applies the ℓ_2-norm over K views (each row of $\mathbf{P}_i$) and the ℓ_1-norm to promote sparsity of $\mathbf{P}_i$, that is $\|\mathbf{P}_i\|_{2,1}=\sum_j \left\|[\mathbf{P}_i]_j\right\|_2$ ($[\mathbf{P}_i]_j$ denotes the j^{th} row of $\mathbf{P}_i$). There are two problems in solving (8): (i) It is based on a non-linear mapping function $\varphi(\bullet)$ on graphs so that the dimension of the feature vector $\varphi(\bullet)$ can be infinitely large, and $\varphi(\bullet)$ may not be explicitly defined. The optimization of (8) is infeasible with any traditional algorithm, such as MOD or KSVD [32]. (ii) The dictionaries for different views $\mathcal{D}_j^k, (k = 1, 2, ..., K)$ are completely independent. The information from multiple views is not effectively combined during dictionary learning.

Our solution to the first problem is inspired by Nguyen et al [33], who proved that the dictionary atoms lie within the subspace spanned by the input training samples. The dictionary $\mathcal{D}_j^k$ can be written as a linear combination of all the training bags as $\mathcal{D}_j^k = \mathbf{V}_j^k \mathbf{S}_j^k, (\mathbf{S}_j^k \in R^{N_j \times M_j})$, where $\mathbf{S}_j^k$ is a linear transformation matrix. It is not necessary to learn the dictionary $\mathcal{D}_j^k$ directly. Instead, we now learn the matrix $\mathbf{S}_j^k$. For the second problem, we set $\mathbf{S}_j^1 = \mathbf{S}_j^2 = ... = \mathbf{S}_j^K = \mathbf{S}_j$ to ensure that the dictionaries from different views share a common transformation matrix $\mathbf{S}_j$. Thus, the objective function (8) is rewritten as

$$\arg\min_{\mathcal{P},\mathbf{S}_j} \sum_{i=1, X_i\in\theta_j}^{N_j} \left\{ \frac{1}{2}\sum_{k=1}^{K} \left\|\varphi(G_{i,k}) - \mathbf{V}_j^k \mathbf{S}_j [\mathbf{P}_i]^k\right\|_2^2 + \gamma\|\mathbf{P}_i\|_{2,1} \right\} \tag{9}$$

In order to balance the sizes of dictionaries from different classes, we set $M_1 = M_2 = ... = M_C = M'$, indicating that the number of atoms in the dictionary for each class is equal to M'. To avoid overfitting, the widely-used penalty regularization on the Frobenius norm of $\mathbf{S}_j \in R^{N_j\times M'}$ with regularization coefficient ξ is added into (9), and the objective function becomes:

$$\begin{aligned}\arg\min_{\mathcal{P},\mathbf{S}_j} \sum_{i=1, X_i\in\theta_j}^{N_j} & \left\{ \tfrac{1}{2}\sum_{k=1}^{K} \left\|\varphi(G_{i,k}) - \mathbf{V}_j^k \mathbf{S}_j [\mathbf{P}_i]^k\right\|_2^2 + \right. \\ & \left. \gamma\|\mathbf{P}_i\|_{2,1} \right\} + \xi\left\|\mathbf{S}_j\right\|_F^2\end{aligned} \tag{10}$$

The optimization of (10) contains two key steps: sparse representation and dictionary update. The implementation of the multi-view dictionary learning is summarized in Algorithm 2 and the details can be found in Appendix B.

The optimization involves the inner products of the vectors of the form $\varphi(\bullet)$ and the inner product in the Reproducing Kernel Hilbert Space (RKHS) can be defined using a kernel function. We use a graph kernel function proposed by Zhou [18]:

$$\begin{aligned} & K_g(G_h, G_q) = \frac{\sum_{a=1}^{n_h}\sum_{b=1}^{n_q} m_{h,a} m_{q,b} Ker(x_{h,a}, x_{q,b})}{\sum_{a=1}^{n_h} m_{h,a}\sum_{b=1}^{n_q} m_{q,b}} \\ & Ker(x_{h,a}, x_{q,b}) = \exp\left(-\kappa\|x_{h,a} - x_{q,b}\|^2\right) \end{aligned}. \tag{11}$$

where $m_{h,a}= 1/\sum_{u=1}^{n_h}\mathbf{M}_h(a,u)$, $m_{q,b}= 1/\sum_{u=1}^{n_q}\mathbf{M}_q(b,u)$; $\mathbf{M}_h$ and $\mathbf{M}_q$ are the adjacency weight matrixes for graphs G_h and G_q, respectively; and $m_{h,a}= 1/\sum_{u=1}^{n_h}\mathbf{M}_h(a,u)$ is a Gaussian radial basis function (RBF) kernel with a parameter κ.

Algorithm 2 Optimization algorithm for multi-view dictionary learning.

- **Input:** the training bags in the j^{th} class, $\theta_j = \{X_i|y_i = j\}$, regularization coefficients γ and ξ, dictionary size M'.
- **Initialize:** Initialize $t = 0$, initialize $[\mathbf{S}_j]_t \in R^{N_j\times M'}$ as a normalized random matrix.
- **repeat**
  - **1) Sparse Representation Step:**
    - For each $X_i \in \theta_j$
    - Compute $\mathbf{P}_i$ by solving (10) with fixed $[\mathbf{S}_j]_t$.
  - **2) Dictionary Update Step:**
    - Set $t = t + 1$;
    - For $k = 1 \rightarrow K$, compute $\mathbf{P}^k = \left[[\mathbf{P}_1]^k, [\mathbf{P}_2]^k, ..., [\mathbf{P}_{N_j}]^k\right] \in R^{M\times N_j}$;
    - Update $[\mathbf{S}_j]_t$ by solving (10).
- **until** convergence of $[\mathbf{S}_j]_t$
- **Output:**$\mathbf{S}_j = [\mathbf{S}_j]_t$.

5 Classification based on M²IL

After obtaining the dictionary $\mathcal{D} = \{\mathcal{D}^1, \mathcal{D}^2, ..., \mathcal{D}^K\}$ using the multi-view dictionary learning, given a test bag with K graphs and an unknown label $(X_T, \Gamma_T =< G_{T,1}, ..., G_{T,K} >, y_T)$, we can obtain the coefficient matrix $\mathbf{W}$ for it by solving (7) with replacing $\varphi(G_{\tau,k})$ with $\varphi(G_{T,k})$. The details of optimization of (7) are given out in Appendix C. Then the reconstruction residual $E_j(X_T)$ of any test bag X_T in class $j \in \{1, 2, ..., C\}$ can be computed as:

$$\begin{aligned} & E_j(X_T) = \sum_{k=1}^{K} \left\|\varphi(G_{T,k}) - \mathcal{D}_j^k \mathbf{W}_j^k\right\|_2^2 \\ & = \sum_{k=1}^{K}\left(1 + [\mathbf{S}_j\mathbf{W}_j^k]^T \mathbf{K}_{\mathbf{V}_j\mathbf{V}_j}^k \mathbf{S}_j \mathbf{W}_j^k - 2[\mathbf{W}_j^k]^T \mathbf{K}_{\mathbf{V}_j\mathbf{T}}^k \mathbf{S}_j\right). \end{aligned} \tag{12}$$

where $\mathbf{K}_{\mathbf{V}_j\mathbf{T}}^k$ is a kernel matrix between the test bag and all the training bags in class j from the k^{th} view, and $\mathbf{K}_{\mathbf{V}_j\mathbf{V}_j}^k$ is a kernel matrix among all the training bags in class j from the k^{th} view. The final class y_T that is assigned to the test bag X_T is the one that gives the smallest reconstruction residual:

$$y_T = \arg\min_{j\in\{1,2,...,C\}} (E_j(X_T)). \tag{13}$$

6 Experiments

This section conducts extensive experiments to evaluate the proposed M^{2}IL algorithm in many practical applications. Further analyses and results are presented in Appendix E.

6.1 Parameter Selection

According to the analysis in Section 3.2, we select 4 parameter settings $\{< \lambda_1, \varepsilon_1 >, < \lambda_2, \varepsilon_2 >, < \lambda_3, \varepsilon_3 >, < \lambda_4, \varepsilon_4 >\} = \{< 1, 0 >, < 0.0001, \varepsilon_2 >, < \lambda_3, 2 >, < \lambda_4, \varepsilon_4 >\}$ corresponding to 4 typical graph structures for each bag. Specially, the graph generated by the proposed sparse ε-graph model with $< \lambda_1, \varepsilon_1 >=< 1, 0 >$ has independent vertices (denoted as "View_1"); the graph with $< \lambda_2, \varepsilon_2 >=< 0.0001, \varepsilon_2 >$ is an ε-graph with the parameter ε_2 (denoted as "View_2"); the graph with $< \lambda_3, \varepsilon_3 >=< \lambda_3, 2 >$ is a ℓ^1-graph with the parameter

TABLE 1
Accuracy (%) on benchmark sets

Algorithm	Musk1	Musk2	Elephant	Fox	Tiger
M^{2}IL	**91.6**($\pm$2.1)	91.0($\pm$1.8)	**89.3**($\pm$0.9)	**65.8**($\pm$1.5)	**87.3**($\pm$1.0)
miGraph	88.9($\pm$3.3)	90.3($\pm$2.6)	86.8($\pm$0.7)	61.6($\pm$2.8)	86.0($\pm$1.6)
MILES	86.3($\pm$3.4)	87.7($\pm$2.8)	86.5($\pm$1.1)	64.7($\pm$2.9)	85.3($\pm$1.0)
MILIS	88.6	**91.1**	N/A	N/A	N/A
MI-SVM	77.9	84.3	81.4	59.4	84.0
mi-SVM	87.4	83.6	82.0	58.2	78.9
missSVM	87.6	80.0	N/A	N/A	N/A
DD	88.0	84.0	N/A	N/A	N/A
EMDD	84.8	84.9	78.3	56.1	72.1

TABLE 2
Accuracy (%) on Image Categorization

Algorithm	1000-Image	2000-Image
M^{2}IL	**84.2**:[82.7,85.3]	**71.2**:[69.5,72.4]
miGraph	82.4:[80.2,82.6]	70.5:[68.7,72.3]
MILIS	83.8:[82.5,85.1]	67.4:[65.3,69.4]
MILES	82.3:[81.4,83.2]	68.7:[67.3,70.1]
MI-SVM	74.7:[74.1,75.3]	54.6:[53.1,56.1]
DD-SVM	81.5:[78.5,84.5]	67.5:[66.1,68.9]
missSVM	78.0:[75.8,80.2]	65.2:[62.0,68.3]
Kmeans-SVM	69.8:[67.9,71.7]	52.3:[51.6,52.9]

λ_3 (denoted as "View_3"); and the final one with $< \lambda_4, \varepsilon_4 >$ is a sparse ε-graph with parameters $0 < \varepsilon_4 < 2$, $\lambda_4 > 0$ (denoted as "View_4"). Therefore, we have 4 parameters $(\varepsilon_2, \lambda_3, \lambda_4, \varepsilon_4)$ in view generation. The other parameters include κ in RBF kernel in (11), regularization coefficient γ in M^{2}IL in (7), and the size of dictionary M'. We set different values of κ in RBF as $\kappa_1, \kappa_2, \kappa_3, \kappa_4$ for different views. We design a greedy parameter selection scheme to determine these parameters efficiently. The details of the scheme are given in Appendix D. According to our experience in the following experiments, the regularization coefficient γ and size of dictionary M' are two key parameters. The value of M' is generally set as around 10% -50% of the average sample number in each class from a data set. The value of γ is generally selected from $\{0.01, 0.1, 1\}$ when the size of data set is medium. Generally, when M' is larger, γ should also be set as a lager value to promote sparsity.

6.2 Experiments on Benchmark Classification Tasks

The first experiment is classification on 5 benchmark data, Musk1, Musk2, Elephant, Fox and Tiger, since they have been extensively used in the studies of MIL. Musk1 contains 47 positive and 45 negative bags, Musk2 contains 39 positive and 63 negative bags, and each of the other three data sets contains 100 positive and 100 negative bags. More details of these five data sets can be found in [1] [14].

We conduct ten-fold cross validations ten times using the procedure described in [18] on these five sets and compare the performance of the M^{2}IL with some leading MIL algorithms, including MI-SVM, mi-SVM [14], MissSVM [17], DD [8], miGraph [18], EM-DD [9], MILIS [7], and MILES [2]. The dictionary size M' is selected from $\{20, 40, 60, 80\}$. The comparisons based on average accuracy and standard deviation values are given in Table 1. The best one for each set is shown in bold. The results of all the other methods are the best results reported in the literature [18], the standard deviations and the results of some algorithms on some sets are not available. The table shows that the M^{2}IL achieves the best performance among all evaluated algorithms on Musk1, Elephant, Fox, and Tiger sets, and comparable performance to MILIS on Musk2. In addition, we notice that the proposed M^{2}IL has lower standard deviations, indicating good stability. From these results, we believe that exploiting context among instances from multiple views can improve the classification accuracy and stability.

6.3 Experiment on Image Classification

The second experiment involves image classification on the COREL image set [3]. The COREL set includes two subsets: COREL-1000 and COREL-2000 that contain 10 and 20 categories of COREL images, respectively. Each category of the two subsets has 100 images. Each image is regarded as a bag, and the regions of interest (ROIs) in the image are regarded as instances described by 9 features [3]. We use the same experimental routine as that described in [3]. For each data set, we randomly partition the images within each category in half, and use one subset for training and leave the other one for testing. The experiment is repeated five times with five random splits, and the average results are recorded. The dictionary size M' in these two sets is selected from $\{20, 40, 60, 80\}$.

The overall accuracy and the 95% confidence intervals are provided in Table 2. For reference, the table also shows the best results of some other MIL methods reported in the literatures, including MI-SVM [14], mi-SVM [14], MissSVM [17], DD-SVM [3], miGraph [18], MILIS [7], MILES [2], and kmeans-SVM [34]. Table 2 shows that the M^{2}IL outperforms all the other algorithms on this set. It shows that integration of multiple views, as in M^{2}IL, is a good method for improving image classification performance.

6.4 Experiment on Image Retrieval

The third experiment evaluates M^{2}ILs performance using an image retrieval task on the SIVAL set created by [35]. The set consists of 25 different objects placed in 10 different scenes. There are 6 different images taken for each object-scene pair, and a total of 1500 images in the set. There is one and only one target object in each image. All the images have been segmented into regions [35]. Each region is represented by a 30D visual feature vector, including the color and texture features, as well as the color and texture differences features [35].

The area under the receiver-operating characteristic (ROC) curve (AUC) [36] [37] is used in this experiment. As in [35], for each category, we use the "one-versus-the-rest" strategy to evaluate the performance. We randomly select 8 positive and 8 negative images to form the training set and let the remaining 1484 images form the test set. The procedure is repeated 30 times with different training samples selections. The dictionary size M' in this set is selected from $\{20, 40, 60\}$.The average AUC values with 95% confidence interval of the 30 rounds of independent tests for the 25 categories are reported in Table3. For comparison, we also list the results of some leading MIL-based CBIR methods, including ACCIO! [35], MILES [2], DD-SVM [3], EC-SVM [36] and MISSL [37]. The performance of the first 4 algorithms is from [36] and the performance of MISSL is from [37].

From the results in Table 3, the performance of the M^{2}IL on 6 categories (FabricSoftenerBox, WD40Can, Coke-Can, FeltFlowerRug, AjaxOrange, and CheckeredScarf) is

TABLE 3
Average AUC values with 95% confidence interval over 30 rounds of test on SIVAL image set.

Category	M^2IL	EC-SVM	MILES	ACCIO!	MISSL	DD-SVM
FabricSoftenerBox	95.0 ± 1.3	**97.9 ± 0.5**	97.1 ±0.7	86.6 ±2.9	97.7 ±0.3	95.7 ±1.8
WD40Can	93.8 ± 1.7	**94.3 ±0.6**	88.1 ± 2.2	82.0 ± 2.4	93.9 ±0.9	86.3 ±2.6
CokeCan	92.8 ± 0.8	**94.6 ± 0.8**	92.4 ± 0.8	81.5 ± 3.4	93.3 ± 0.9	94.0 ± 0.9
FeltFlowerRug	93.4 ± 1.0	**94.2 ± 0.8**	93.9 ± 0.7	86.9 ± 1.6	90.5 ± 1.1	91.4 ± 0.7
AjaxOrange	93.7 ± 2.3	**93.8 ± 2.1**	90.2 ± 2.3	77.0 ± 3.4	90.0 ± 2.1	84.1 ± 3.2
CheckeredScarf	**94.6 ± 0.7**	96.9 ± 0.5	93.7 ± 1.2	90.8 ± 1.5	88.9 ± 0.7	96.2 ± 0.7
CandleWithHolder	**89.7 ± 0.9**	88.1 ± 1.1	84.0 ± 2.3	68.8 ± 2.3	84.5 ± 0.8	77.3 ± 2.8
GoldMedal	**88.9 ± 1.2**	87.5 ± 1.4	80.7 ± 2.9	77.7 ± 2.6	83.4 ± 2.7	73.4 ± 4.1
SpriteCan	**87.7 ± 1.5**	85.4 ±1.2	80.4 ±2.0	71.9 ± 2.5	81.2 ± 1.5	81.1 ± 2.4
SmileyFaceDoll	**88.3 ±2.1**	84.6± 1.9	77.5± 2.6	77.4± 3.3	80.7± 2.0	69.3± 3.9
GreenTeaBox	**89.4 ± 2.3**	86.9 ± 2.2	91.2 ± 1.7	87.3 ± 3.0	80.4 ± 3.5	86.9 ± 3.1
DirtyRunningShoe	**91.4 ± 1.8**	90.3 ± 1.3	85.3 ± 1.7	83.7 ± 1.9	78.2 ± 1.6	87.3 ± 1.4
DataMiningBook	**79.4 ± 2.6**	75.0 ± 2.4	71.1 ± 3.2	74.7 ± 3.4	77.3 ± 4.3	68.8 ± 3.7
BlueScrunge	**78.3 ± 2.1**	74.1 ± 2.4	72.6 ± 2.5	69.5 ± 3.4	76.8 ± 5.2	62.1 ± 2.9
DirtyWorkGloves	**85.1 ± 1.7**	83.0 ± 1.3	77.1 ± 3.1	65.3 ± 1.5	73.8 ± 3.4	72.3 ± 2.2
StripedNotebook	**78.9±2.0**	75.6±2.3	68.7± 2.4	70.2± 3.2	70.2± 2.9	67.3± 3.0
CardboardBox	**87.3± 1.9**	85.6± 1.6	81.2± 2.7	67.9± 2.2	69.6± 2.5	73.0± 3.0
JuliesPot	**84.3± 2.3**	67.3± 3.3	78.7± 2.9	79.2± 2.6	68.0± 5.2	74.3± 3.0
TranslucentBowl	**79.9± 2.5**	74.2± 3.2	73.2± 3.1	77.5± 2.3	63.2± 5.2	67.3± 2.7
Banana	**74.2± 2.7**	69.1± 2.9	68.1± 3.1	65.9± 3.3	62.4± 4.3	62.2± 1.6
RapBook	**76.2± 2.1**	68.6± 2.3	61.7± 2.4	62.8± 1.7	61.3± 2.8	66.2± 2.0
WoodRollingPin	**73.4± 1.9**	66.9± 1.7	62.1± 2.5	66.7± 1.7	51.6± 2.6	64.8± 1.4
GlazedWoodPot	**76.4± 2.4**	68.0± 2.8	68.2± 3.1	72.7± 2.3	51.5± 3.3	68.2± 3.4
Apple	**76.9± 3.1**	68.0± 2.6	64.5± 2.5	63.4± 3.4	51.1± 4.4	62.8± 2.3
LargeSpoon	**75.8± 1.7**	61.3± 1.8	58.2± 1.6	57.6± 2.3	50.2± 2.1	59.7± 1.8
Average	**85.0**	81.3	78.4	74.6	74.8	75.7

slightly lower than that of the EC-SVM method. However, the M^2IL outperforms the existing methods with obvious performance improvements on the other 19 categories. The higher AUC values (larger than 90) of EC-SVM on the 6 categories indicate that the retrieval task on these 6 categories is relatively easier than the other categories. Besides obtaining comparable performance on the easy categories, the proposed M^2IL also achieve much better performance on the difficult categories. This performance improvement can be ascribed to the integration of the multi-view cues in the categories.

TABLE 4
Performance (%) on Horror Video Recognition

Algorithm	Precision	Recall	F-measure
M^2IL	**88.6(±0.43)**	**87.8(±0.39)**	**88.2(±0.41)**
miGraph	81.8(±1.95)	82.4(±1.25)	82.1(±1.2)
MI-kernel	80.7(±1.42)	81.4(±0.9)	81.1(±0.5)
MI-SVM	79.8	78.9	79.4
mi-SVM	75.4	75.4	75.4
CKNN	78.9	70.5	74.5
EM-DD	77.6	73.0	75.2

6.5 Experiment on Horror Video Recognition

The final experiment is to test the M^2IL on a video recognition task. We investigate this task using the M^2IL on a horror video set [38]. This set contains 400 horror movie scenes and 400 non-horror movie scenes in total. Each movie scene is viewed as a "bag" and divided into a series of shots via shot detection. The key frame of each shot is extracted as an "instance" in the bag. Each frame is represented as a 153D feature vector, including color feature, audio feature, and affective feature [38].

The dictionary size M' on this set is selected from $\{100, 200, 300\}$. For each algorithm, the average precision, recall, F-measure [38], and corresponding standard deviation values of ten times 10-fold cross validation are used as the final performance as shown in Table 4. The results of the MI-SVM, mi-SVM [14], CKNN [11], EM-DD [9] in Table 4 are from [38] by Wang et al. The standard deviations of these 4 algorithms are not reported in [38].

The results in Table 4 show that the M^2IL and miGraph methods outperform the other methods, which further indicates that the context is useful in horror video recognition. The fact that performance of M^2IL is much better than the performance of miGraph and MI-kernel shows that horror scene recognition can benefit from considering context from multiple views. According to this experiment, we can find that the multi-view contextual structures embedded in the M^2IL can effectively express the relations among frames in a video.

6.6 Single View vs. Multiple Views

We use 4 views in the M^2IL (denoted as "View_All" here) in previous experiments. The M^2IL can also use with only one view. We test such single view-based M^2IL methods using "View_1", "View_2", "View_3", and "View_4", respectively. The experimental settings and routines of these four single view-based methods are the same as the M^2IL with all views and the comparison results are shown in Figure 3. We can find that: (i) No single view consistently achieves obviously better performance than the others. It again indicates that it is difficult to well represent the relations among instances in a bag using a fixed structure for different tasks. (ii) The M^2IL integrating all views always outperforms the others, showing that considering multiple views can effectively improve the performance of MIL.

7 Conclusion

This paper proposes a multi-view multi-instance learning (M^2IL) algorithm where the "multi-view" is defined as a series of graphs to represent the inherent contextual structures among instances in a bag. We propose a sparse ε-graph model that can generate multiple undirected graphs

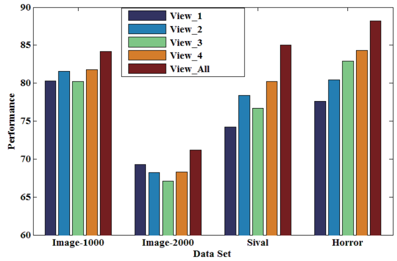

Fig. 3. The comparison of the M^2IL with different views on four data sets.

for different parameter values to represent different inner contextual structures among instances in a bag. Then all of these contextual structures are simultaneously considered under a proposed multi-view joint sparse representation framework for bag classification. A novel multi-view dictionary learning framework is also presented to improve the performance and robustness of the M^2IL. Experimental results and analyses show that integrating multiple inner contextual structures from different views can improves the performance of MIL.

ACKNOWLEDGEMENTS

This work is partly supported by the Natural Science Foundation of China (Grant Nos. 61370038, U1636218, 61472421, 61571045), the 973 basic research program of China (Grant No. 2014CB349303), the Strategic Priority Research Program of the CAS (Grant No. XDB02070003), and the CAS External cooperation key project. Bing Li is also supported by Youth Innovation Promotion Association, CAS.

REFERENCES

[1] T. G. Dietterich, R. H. Lathrop and T. Lozano-Perez, Solving the multiple-instance problem with axis-parallel rectangles, Artif. Intell., vol. 89, no. 1-2, pp. 31-71, 1997.
[2] Y. Chen, J. Bi, and J. Z. Wang, MILES: Multiple-instance learning via embedded instance selection, IEEE Trans. Pattern Anal. Mach. Intell., vol. 28, no. 12, pp. 1931-1947, 2006.
[3] Y. Chen, and J. Wang, Image categorization by learning and reasoning with regions, J. Mach. Learn. Res., vol.5, pp. 913-939, 2004.
[4] Q. Zhang, W. Yu, S. A. Goldman, and J. E. Fritts, Content-based image retrieval using multiple-instance learning, Proc. Intl Conf. Machine Learning, pp. 682 - 689, 2002.
[5] J. Amores, Multiple instance classification: Review, taxonomy and comparative study, Artif. Intell., vol. 201, pp. 81-105, 2013.
[6] J. Foulds, and E. Frank, A review of multi-instance learning assumptions, Knowl. Eng. Rev., vol. 25, no. 1, pp. 1-25, 2010.
[7] Z. Fu, A. Robles-Kelly, and J. Zhou, MILIS: Multiple instance learning with instance selection, IEEE Trans. Pattern Anal. Mach. Intell., vol. 33, no. 5, pp. 958-977, 2011.
[8] O. Maron and T. Lozano-Perez, A Framework for Multiple Instance Learning, Proc. Conf. Advances in Neural Information Processing Systems, pp. 570-576, 1998.
[9] Q. Zhang and S. Goldman, Em-DD: An Improved Multiple Instance Learning Technique, Proc. Conf. Advances in Neural Information Processing Systems, pp. 1073-1080, 2002
[10] R. Rahmani, S.A. Goldman, H. Zhang, S.R. Cholleti, and J.E. Fritts, Localized Content Based Image Retrieval, Trans. Pattern Anal. Mach. Intell., vol. 30, no. 11, pp. 1902-2002, 2008.
[11] J. Wang and J.D. Zucker, Solving the Multiple-Instance Problem: A Lazy Learning Approach, Proc. Intl Conf. Machine Learning, pp. 1119-1125, 2000.
[12] H. Wang, H. Huang, F. Kamangar, F. Nie, and C. Ding. Maximum Margin Multi-Instance Learning, Proc. Conf. Advances in Neural Information Processing Systems, pp. 1-9, 2011.
[13] T. Gartner, A. Flach, A. Kowalczyk, and A.J. Smola, Multi-Instance Kernels, Proc. Intl Conf. Machine Learning, pp. 179-186, 2002.
[14] S. Andrews, I. Tsochantaridis, and T. Hofmann, Support vector machines for multiple instance learning, Proc. Conf. Advances in Neural Information Processing Systems, pp. 561 - 568, 2003.
[15] C. Bergeron, G. Moore, J. Zaretzki, C. M. Breneman, and K. P. Bennett, Fast Bundle Algorithm for Multiple-Instance Learning, Trans. Pattern Anal. Mach. Intell.,vol. 34, no. 6, 1068-1077, 2012.
[16] P.-M. Cheung and J.T. Kwok, A Regularization Framework for Multiple-Instance Learning, Proc. Intl Conf. Machine Learning, pp. 193-200, 2006.
[17] Z. H. Zhou and J. M. Xu, On the Relation between Multi-Instance Learning and Semi-Supervised Learning, Proc. Intl Conf. Machine Learning, pp. 1167-1174, 2007.
[18] Z. H. Zhou, Y. Sun, and Y. Li, Multi-Instance Learning by Treating Instances As Non-I.I.D. Samples, Proc. Intl Conf. Machine Learning, pp. 1249-1256, 2009.
[19] J. B. Tenenbaum, V. de Silva, and J. C. Langford, A global geometric framework for nonlinear dimensionality reduction, Science, vol. 290, 2319C2323, 2000.
[20] H. J. Song, J. W. Son, and S. B. Park, Identifying User Attributes through non-i.i.d. Multi-Instance Learning, Proc. IEEE/ACM Conf. Advances in Social Networks Analysis and Mining, pp. 25-28, 2013.
[21] B. Li, W.H. Xiong, O. Wu, and W. M. Hu, horror image recognition based on context-aware multi-instance learning, IEEE Trans. Image Process., vol. 24, no. 12, pp. 5193-5025, 2015.
[22] X. Ding, B. Li, and W. Hu, Horror Video Scene Recognition based on Multi-view Multi-instance Learning, Proc. Asian Conf. Computer Vision, PP.599-610, 2012.
[23] D. Zhang, Y. Liu, L. Si, J. Zhang, and R. D. Lawrence, Multiple Instance Learning on Structured Data, Proc. Conf. Advances in Neural Information Processing Systems, pp. 145-153, 2011.
[24] C. Xu, D. Tao, and C. Xu, A Survey on Multi-view Learning, CoRR abs/1304.5634, 2013.
[25] B. Cheng, J. Yang, S. Yan, Y. Fu, and T. Huang, Learning with L1-Graph for Image Analysis, IEEE Trans. on Image Processing, vol. 19, no. 4, pp. 858-866, 2010.
[26] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong, Locality-constrained Linear Coding for Image Classification, Proc. IEEE Conf. Comput.Vis. Pattern Recog., pp. 1063-6919, 2010.
[27] J. Wright, Y. Ma, J. Mairal, G. Sapiro, Sparse Representation for Computer Vision and Pattern Recognition. Proceedings of the IEEE, vol. 98, no. 6, pp. 1031-1044, 2010.
[28] J. Wright, A. Yang, A. Ganesh, S. Sastry, and Y. Ma. Robust face recognition via sparse representation. IEEE Trans. Pattern Anal. Mach. Intell., vol. 31, no. 2, pp. 210-227, 2009.
[29] X. Yuan, X. Liu, and S. Yan, Visual Classification With Multitask Joint Sparse Representation, IEEE Trans. Image Process.,, vol. 21, no. 10, pp. 4349-4360, 2012.
[30] E. Elhamifar, G. Sapiro, and R. Vidal. See All by Looking at A Few: Sparse Modeling for Finding Representative Objects, Proc. IEEE Conf. Comput.Vis. Pattern Recog., pp. 1600-1607, 2012.
[31] M. Yang, L. Zhang, J. Yang, and D. Zhang. metaface learning for sparse representation based face recognition. Proc. IEEE Int. Conf. on Image Processing, pp. 1601-1604, 2010.
[32] M. Aharon, M. Elad, and A. M. Bruckstein, The K-SVD: An algorithm for designing of overcomplete dictionaries for sparse representation, IEEE Trans. Signal Process., vol. 54, no. 11, pp. 4311-4322, 2006.
[33] H. V. Nguye, V. M. Patel, N. M. Nasrabadi, and R. Chellappa, Design of Non-Linear Kernel Dictionaries for Object Recognition, IEEE Trans. Image Process.,, vol. 22, no. 12, pp. 5123-5135, 2013.
[34] G. Csurka, C. Bray, C. Dance, and L. Fan, Visual categorization with bags of keypoints, Proc. ECCV Workshop on Statistical Learning in Computer Vision, pp. 59-74, 2004.
[35] R. Rahmani, S. A. Goldman, H. Zhang, J. Krettek, and J. E. Fritts, Localized content based image retrieval, Proc. ACM International Conference on Multimedia Information Retrieval, pp.227-236, 2005.
[36] W. Li, and D. Yeung, Localized content-based image retrieval through evidence region identification, Proc. IEEE Conf. Comput.Vis. Pattern Recog., pp. 1666-1673, 2009.
[37] R. Rahmani, and S. A.Goldman, Missl: multiple-instance semi-supervised learning, Proc. Intl Conf. Machine Learning, pp. 705-712, 2006.
[38] J. C. Wang, B. Li, and W. M. Hu, Horror video scene recognition via mutiple-instance learning, Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., pp. 1325-1328, 2011.