ABSTRACT
Object-based image retrieval has been an active research topic in recent years, in which a user is only interested in some object in the images. As one promising approach, graph-based multi-instance learning has attracted many researchers. The existing methods often conduct learning on one graph, either in image level or in region level. While in this paper, by considering both image- and region-level information at the same time, a novel method based on multi-graph multi-instance learning is proposed. Two graphs are constructed in our method, and the relationship between each image and its segmented regions is introduced into an optimization framework. Moreover, our method is further extended to video retrieval. By exploring the relationships between video shots, representative images, and segmented regions, it can deal with the case when training labels are only assigned in shot level. Experimental results on the SIVAL image benchmark and the TRECVID video set demonstrate the effectiveness of our proposal.

Categories and Subject Descriptors
H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval—retrieval models

General Terms
Algorithms, Experimentation

Keywords
Object-based retrieval, graph-based learning, multi-instance learning

1. INTRODUCTION
With the explosive growth of the number of digital images and videos, how to well organize and manage media data has attracted much attention, and effective retrieval technique is in urgent need. Because of its large labeling cost, text-based methods are obviously impractical. Therefore, content-based image and video retrieval, which is based on automatically extracted features, has become an active research topic in the last decades [5], [16], [21], [22], [31].

When conducting visual information retrieval, a user usually does not pay attention to the whole image or video. Instead, the user may be only interested in the portion corresponding to some object. As a result, if only global features are adopted, the performance often cannot satisfy the user’s need. To deal with the problem, object-based (or localized content-based) retrieval is proposed, and much related work has been developed [12], [18], [20], [33], [34], [35].

As an effective approach to describe the relationship between whole and part, multi-instance learning, which is originally proposed for drug activity prediction [6], has been widely used in the field of media analysis and understanding [1], [2], [3], [8], [13], [32]. In the framework of multi-instance learning, each sample is called a bag, and it contains several instances. The available label information is only assigned for bags, and the relationship between bag and instance is that a bag is positive if at least one instance in it is positive, otherwise it is negative. In order to involve object-based image retrieval into the framework of multi-instance learning, images are first segmented into regions, and then images and regions are treated as bags and instances, respectively. In the early stage, most multi-instance methods only learn from labeled images, and conduct retrieval in supervised manners. For effectively exploring data relationship and making full use of unlabeled images to further improve performance, semi-supervised multi-instance image retrieval, especially graph-based learning, has attracted more and more attention in recent years.

According to the constructed graph, the existing graph-based multi-instance image retrieval methods can be generally divided into two categories. In the first category, each vertex in the graph corresponds to one image, and multi-instance learning is ultimately transformed into image-level single-instance learning. In order to introduce region-level information, one approach is to calculate the weighted edges of image-level graph based on region similarities [19], and another approach is to map all the images into a new space spanned by the selected prototypes of regions [25]. Regardless which approach is adopted, since there are often several foreground objects as well as cluttered background in each image, using only one weight cannot well reflect the relationship between two images.

In the other category of graph-based multi-instance image retrieval methods, the graph is constructed in region level.
In [24], some underlying positive regions are first selected, and then propagate their real-valued labels to the regions of all the database images. In the process of region selection, it is implied that at least one positive image only includes one underlying positive region. To avoid such an additional restriction, the idea of weighting regions in positive images is introduced in [17]. After analyzing the learning results on all the regions in training images, higher weights are assigned to the regions corresponding to the user-interested object. Some other methods consider the relationship between each image and its segmented regions, as well as the available image-level training labels, as the fitting constraint in a cost function, and calculate region soft labels by solving an optimization problem [14], [27]. Since the characteristics of both graph-based learning and multi-instance learning are well involved in a unified framework, these methods can achieve the state-of-the-art performance.

In most of the existing graph-based multi-instance methods, no matter which kind of graph is constructed, image-level features are totally discarded. Although global color and texture cannot effectively describe the user-interested object, the idea of “bag of key points” [4] can still be utilized. Since key point features can also represent the characteristics of each object, their image-level statistical description is also useful for visual categorization and retrieval [7], [9].

In this paper, based on the idea of combining both image- and region-level information together, we present a novel graph-based multi-instance learning framework. Two graphs are adopted in our proposal, one is constructed by bag-of-key-points features, and the other is constructed by region features. Since their sizes are different, the existing graph-based multi-modality learning methods [26], [29] cannot be utilized. To deal with the problem, the whole learning process is conducted in an optimization framework. And the relationship between images, the relationship between regions, as well as the relationship between each image and its segmented regions, are all well involved. Since key point features and region features describe the user-interested object from different views, the retrieval performance can be improved by considering the two kinds of descriptions at the same time.

Moreover, we extend our proposal to video retrieval. When the available label information is assigned for each representative image selected in the training shots, the methods for image retrieval can be directly adopted for searching videos. However, if we only know whether or not each training shot is relevant to the user’s query, most of the existing image retrieval methods cannot deal with the case. To address the problem, multi-layer multi-instance learning is well studied in [10]. However, its regularization framework only involves classification error and inter-layer inconsistency, while the data relationship in each layer is not considered. With the help of graph-based representation, our extended video retrieval method can take all the aforementioned factors into account, and conduct effective learning in a unified optimization framework.

The rest of the paper is organized as follows. Section 2 describes our proposed multi-graph multi-instance learning method for image retrieval. Then the method is extended to video retrieval in Section 3. Section 4 presents our experimental results on image and video data sets, and it is followed by some conclusions and analysis of future work in Section 5.

2. MULTI-GRAph Multi-InStANCE LEARNING FOR IMAGE RETRIEVAL

In this section, first we explain the ways to represent images and to construct weighted graphs. Then we present the optimization framework and talk about its solution. At last, we discuss some extensions of our proposal.

2.1 Image Representation

Suppose there are altogether $M$ images denoted as $I = \{I_1, I_2, \ldots, I_M\}$, which can be divided into three sets, the positive image set $I^+$, the negative image set $I^-$, and the unlabeled image set $I^0$. After segmentation, each image $I_m (m = 1, 2, \ldots, M)$ is represented by a set of regions. Let the total number of regions be $N$. When the corresponding images are unnecessary to point out, the regions are denoted as $\mathcal{R} = \{R_1, R_2, \ldots, R_N\}$. And we use $R_n \in I_m$ to indicate that $R_n$ is a region in the image $I_m$. The sets of regions in the positive images, in the negative images, and in the unlabeled images are denoted as $\mathcal{R}^+, \mathcal{R}^-$, and $\mathcal{R}^0$, respectively. It should be noted that because of the multi-instance relationship, not all the regions in $\mathcal{R}^+$ are relevant to the user’s query.

2.2 Graph Construction

To effectively represent the relationship between images and that between regions, two weighted graphs are constructed. The image-level graph includes $M$ vertices, each of which corresponds to an image. For avoiding dense graph and reducing computational load, the neighbor set of each image $I_m$, denoted as $\mathcal{N}(I_m)$, is found, and the element in the affinity matrix $W^I$ is given by

$$w^I_{mn} = \begin{cases} \exp\left(-\frac{\|p(I_m) - p(I_n)\|^2}{2\sigma^2}\right), & \text{if } I_m \in \mathcal{N}(I_n) \\ 0, & \text{otherwise.} \end{cases}$$

(1)

where $p(I_m)$ and $p(I_n)$ are key point features of $I_m$ and $I_n$, $\sigma$ is a parameter to control the shape of Gaussian function.

Similarly, the region-level graph with $N$ vertices is constructed, and the element in the affinity matrix $W^R$ is also calculated by Gaussian function. In order to avoid self-reinforcement and improper propagation, when finding the neighbor set of each region, the regions in the same image are not taken into consideration.

2.3 Optimization Framework

For determining the final retrieval results, two real-valued vectors, namely image soft labels and region soft labels, are introduced, and they are denoted as $f^I = [f^I_1, f^I_2, \ldots, f^I_M]^T$ and $f^R = [f^R_1, f^R_2, \ldots, f^R_N]^T$, respectively. It is hoped that the values of soft labels can reflect the relevance to the user’s query. The larger the soft label is, the more likely the corresponding image or region satisfies the user’s need. In order to find the optimal vectors, a cost function $Q(f^I, f^R)$ should be designed, and the soft labels will be calculated by solving the optimization problem to minimize the cost function. The framework of learning with local and global consistency [36] is adopted in our proposal, in which both smoothness and fitting constraints are involved to design the cost function.

With the basic idea that the soft labels of similar images and regions should not change too much, the cost terms corresponding to the smoothness constraint in the two weighted
graphs are defined in the same form as those used in the existing graph-based learning methods [11], [30]

\[ Q_{SMT}(t^I) = \frac{1}{2} \sum_{m,n=1}^{M} W^I_{m,n} \left( \frac{f^I_m - f^I_n}{\sqrt{d_m}} \right)^2 \]  

(2)

\[ Q_{SMT}(t^R) = \frac{1}{2} \sum_{m,n=1}^{N} W^R_{m,n} \left( \frac{f^R_m - f^R_n}{\sqrt{d_n}} \right)^2 \]  

(3)

where \(d_m^I\) and \(d_n^R\) equal to the sum of the \(m\)-th row of \(W^I\) and \(W^R\) respectively.

In graph-based learning, the fitting constraint often means that good soft labels should not change too much from the initial label assignment. Since the available label information is only provided in image level, and the user-interested regions are not explicitly indicated, different approaches are used to design the cost items corresponding images and regions.

For labeled images, as we have known whether or not they are relevant to the user’s query, the cost items are defined to calculate the difference between the image soft labels and the corresponding initial labels

\[ Q_{FIT}^+(t^I) = \sum_{I_m \in \mathcal{I}^+} H_1(f^I_m, 1) \]  

(4)

\[ Q_{FIT}^-(t^I) = \sum_{I_m \in \mathcal{I}^-} H_2(f^I_m, -1) \]  

(5)

where \(H_1(\cdot, \cdot)\) and \(H_2(\cdot, \cdot)\) are suitable distance measures. Euclidean distance is usually adopted, but here we only want to punish the positive images whose soft labels are smaller than 1, as well as the negative images whose soft labels are larger than -1, so the measures are defined as

\[ H_1(x, y) = (\max(y - x, 0))^2 \]  

(6)

\[ H_2(x, y) = (\max(x - y, 0))^2 \]  

(7)

In this case, the positive images with very large soft labels and the negative images with very small soft labels will not have any influence on the values of the cost items.

For regions in labeled images, they are not assigned with initial labels, hence the fitting constraint is determined by analyzing the relationship between each image and its segmented regions. According to the characteristics of multi-instance learning, the approaches for dealing with regions in positive and negative images are different. As far as regions in positive images are concerned, since we do not know which region is relevant to the user’s query, only the region with the largest soft label is taken into consideration. In this paper, the cost item is defined to calculate the difference between the largest region soft label and the image soft label, namely

\[ Q_{FIT}^+(t^R) = \sum_{I_m \in \mathcal{I}^+} H_1\left(\max_{R_n \in \mathcal{R}_m} f^R_n, f^I_m\right) \]  

(8)

As all the regions in negative images should also be irrelevant to the user’s query, the cost item is simply defined as

\[ Q_{FIT}^-(t^R) = \sum_{R_n \in \mathcal{R}^-} H_2(f^R_n, -1) \]  

(9)

At last, the smoothness and fitting constraints are combined together, and the final cost function is defined as

\[ Q(t^I, t^R) = \mu_G^I Q_{SMT}(t^I) + \mu_G^R Q_{SMT}(t^R) + \mu^I Q_{FIT}^+(t^I) + \mu^R Q_{FIT}^-(t^I) \]  

(10)

where \(\mu_G^I, \mu_G^R, \mu^I, \mu^R, \mu^L, \mu^R\) are combination coefficients. To reduce the number of parameters, some prior knowledge can be introduced. As positive images share the same object and negative images are irrelevant from different ways, the cost terms corresponding to positive images should be enhanced. Moreover, since the user’s query is provided in image level, and the segmentation results are not always accurate enough, we should pay more attention to image-level information. Therefore, the combination coefficients can be set as \(\mu_G^I = \delta (\delta \geq 1), \mu_G^R = 1, \mu^I = 2\mu^L, \mu^R = 3\mu^R (\beta \geq 1)\), and \(\mu^L = \alpha (\alpha \geq 0)\), where \(\alpha, \beta, \delta\), and\( \delta\) are parameters to be predefined.

### 2.4 Solution to Optimization Problem

To solve the optimization problem, first the designed cost function (10) is rewritten as

\[ Q(t^I, t^R) = \frac{\mu_G^I}{2} \sum_{m,n=1}^{M} W^I_{m,n} \left( \frac{f^I_m - f^I_n}{\sqrt{d_m}} \right)^2 \]  

\[ + \frac{\mu_G^R}{2} \sum_{m,n=1}^{N} W^R_{m,n} \left( \frac{f^R_m - f^R_n}{\sqrt{d_n}} \right)^2 \]  

\[ + \mu^I \sum_{I_m \in \mathcal{I}^+} \xi^2_m + \mu^R \sum_{I_m \in \mathcal{I}^-} \eta^2_m \]  

\[ + \mu^L \sum_{R_n \in \mathcal{R}^-} \zeta^2_n \]  

(11)

with the constraints

\[ \begin{align*}
1 - f^I_m &\leq \xi_m, \quad \xi_m \geq 0, \quad (I_m \in \mathcal{I}^+); \\
1 + f^I_m &\leq \lambda_m, \quad \lambda_m \geq 0, \quad (I_m \in \mathcal{I}^-); \\
 f^R_m - \max_{R_n \in \mathcal{R}_m} f^R_n &\leq \eta_m, \quad \eta_m \geq 0, \quad (I_m \in \mathcal{I}^+); \\
1 + f^R_m &\leq \zeta_n, \quad \zeta_n \geq 0, \quad (R_n \in \mathcal{R}^-).
\end{align*} \]  

(12)

For dealing with the optimization problem like this, the constrained concave convex procedure (CCCP) [23] can be adopted. It is an iterative process, and the 1st-order Taylor expansions are used to approximate the non-convex functions in each step. After iteratively solving the approximated convex optimization problem, the sub-optimal solution can be found. To deal with the problem that the max(·) function is not differentiable at all points, its sub-gradient is adopted. For the function of max(\(\mathbf{x}\)), the elements of its sub-gradient \(\Delta\) can be calculated as

\[ \Delta_i = \begin{cases}
1/k, & \text{if } x_i = \max(\mathbf{x}); \\
0, & \text{otherwise}.
\end{cases} \]  

(13)

where \(k\) is the number of elements whose values equal to the maximum of \(\mathbf{x}\). For more details about the solution to such an optimization problem, please refer to [14], [37].

After the soft labels for images and regions are calculated, the final relevance score \(\phi_m^I\) for image \(I_m\) is determined by combining the two kinds of information together

\[ \phi_m^I = \omega_1 f^I_m + (1 - \omega_1) \max_{R_n \in \mathcal{R}_m} f^R_n \]  

(14)
where \( \omega_i (0 \leq \omega_i \leq 1) \) is a tunable parameter reflecting our confidence on the two kinds of soft labels, and can simply set to 0.5 for convenience. Then the images with the largest relevance scores are returned as the retrieval results.

### 2.5 Extensions

When there are no unlabeled images for training, our proposal can be reduced to a supervised method. Since unlabeled images and their segmented regions are only involved in the smoothness constraint for semi-supervised learning, they can be easily removed from the weighted graphs. Therefore, without unlabeled training images, the cost terms are in the same form as (2) and (3), the only difference is that \( M \) and \( N \) denote the total numbers of labeled images and their segmented regions.

To deal with out-of-sample images which are not available in the training stage, the same smoothness constraint can be easily removed from the weighted graphs. Thereby, the idea of multi-instance learning, all the images and regions in the same form as (2) and (3). Moreover, due to the basic idea of multi-instance learning, all the images and regions in negative shots are also irrelevant to the user’s query. Therefore, the definitions for the terms of \( Q_{\text{FIT}}(f^I) \) and \( Q_{\text{FIT}}(f^R) \) can also be the same as (5) and (9).

When training labels are only assigned in shot level. First, some criterion is set to apply the same smoothness constraint for semi-supervised learning, the only difference is that the total numbers of labeled images and their segmented regions are only involved in the smoothness constraint for semi-supervised learning, the terms of \( Q_{\text{SMT}}(f^I) \) and \( Q_{\text{SMT}}(f^R) \) can be defined in the same way as (2) and (3). Moreover, due to the basic idea of multi-instance learning, all the images and regions in negative shots are also irrelevant to the user’s query. Therefore, the definitions for the terms of \( Q_{\text{FIT}}(f^I) \) and \( Q_{\text{FIT}}(f^R) \) can also be the same as (5) and (9).

### 3. MULTI-GRAF MULTI-INSTANCE LEARNING FOR VIDEO RETRIEVAL

In this section, we consider the case for video retrieval when training labels are only assigned in shot level. First, we introduce some notations for representing video shots. Then we discuss the extended optimization framework and its solution in detail.

#### 3.1 Shot Representation

Let the collection of video shots be denoted as \( S = \{S_1, S_2, \cdots, S_L\} \), where \( L \) is the shot number. And the sets of positive, negative and unlabeled shots are denoted as \( S^+, S^- \), and \( S^U \), respectively. For shot representation, first some representative images are chosen, which can be conducted by either simply sampling or elaborately selecting based on some criterion. Then each representative image is further segmented into regions. For convenience, the same notations as those in Section 2 are used, and the sets of representative images and segmented regions are denoted as \( I = \{I_1, I_2, \cdots, I_M\} \) and \( R = \{R_1, R_2, \cdots, R_N\} \), respectively. We also use \( I_m \in S_l, R_n \in S_l \) and \( R_n \in S_l \) to indicate the relationships between shot, image and region. Besides, the notations of \( I^+, I^- \), and \( I^U \), as well as \( R^+, R^- \), and \( R^U \), are also adopted. It should be noted that \( I^+ \) denotes all the representative images in positive shots, and not all of them are positive. The situation is the same for \( R^+ \).

#### 3.2 Extended Optimization Framework

After the image- and region-level graphs are constructed by the method in Section 2.2, we also introduce image soft labels \( f^I_m \ (m = 1, 2, \cdots, M) \) and region soft labels \( f^R_n \ (n = 1, 2, \cdots, N) \), whose values will be calculated by solving another optimization problem. With the similar idea for taking both smoothness and fitting constraints into account, the final cost function should have the same form as (10).

As smoothness constraint only considers the relationship between images or regions, and no label information is involved, the terms of \( Q_{\text{SMT}}(f^I) \) and \( Q_{\text{SMT}}(f^R) \) can be defined in the same way as (2) and (3). Moreover, due to the basic idea of multi-instance learning, all the images and regions in negative shots are also irrelevant to the user’s query. Therefore, the definitions for the terms of \( Q_{\text{FIT}}(f^I) \) and \( Q_{\text{FIT}}(f^R) \) can also be the same as (5) and (9).

When training labels are only assigned for video shots, it is unknown whether or not the representative images in each positive shot are relevant to the user’s query. As a result, the equations of (4) and (8) cannot be used any more, and the terms of \( Q_{\text{FIT}}(f^I) \) and \( Q_{\text{FIT}}(f^R) \) should be redefined. As far as representative images in each positive shot are concerned, we only consider the image with the largest soft label, and define the cost item as

\[
Q_{\text{FIT}}(f^I) = \sum_{S_t \in S^+} H_1 \left( \max_{I_m \in S_t} f^I_m, 1 \right) \tag{17}
\]

For regions in positive shots, the case is more complicated. As there is an image-layer between regions and shots, the image-level information should also be taken into consideration. If it has been known which image is relevant to the user’s query, for regions in each positive image, the cost term can also be defined as the difference between the largest region soft label and the image soft label. Thus a set of possible positive images, denoted as \( C^+ \), is introduced, and the cost item is defined as

\[
Q^+_{\text{FIT}}(f^R) = \sum_{I_m \in C^+} H_1 \left( \max_{R_n \in I_m} f^R_n, f^I_m \right) \tag{18}
\]

By combining the aforementioned six terms, the final cost function about \( f^I \) and \( f^R \) can be defined. The problem left is how to determine the set of \( C^+ \), we will discuss it in the next section.

#### 3.3 Solution to Optimization Problem

In this paper, the above problem is treated as joint optimization for \( f^I \) and \( f^R \). In order to solve it, an iterative approach is proposed, in which \( f^I \) and \( f^R \) are calculated respectively, and the details of the process are explained below.

At first, the original values of elements in \( f^I \) are defined as

\[
f^I_m = \begin{cases} 
1, & \text{if } I_m \in I^+; \\
-1, & \text{if } I_m \in I^-; \\
0, & \text{if } I_m \in I^U. 
\end{cases} \tag{19}
\]
With fixed \( f^l \), the original cost function is reduced to

\[
Q(f^R) = \frac{\mu_R^2}{2} \sum_{m,n=1}^{N} W_{mn} \left( \frac{f^R_m}{\sqrt{d^R_m}} - \frac{f^R_n}{\sqrt{d^R_n}} \right)^2 + \mu_R \sum_{l_m \in C^+} H_1 \left( \max_{l_n \in l_m} f^R_n, f^l_m \right) + \mu_R \sum_{l_n \in R^+} H_2 \left( f^R_n, -1 \right)
\]  

(20)

At this time, \( C^+ \) is determined by comparing the values of \( f^R_m \) and a threshold \( TH^I \)

\[
C^+ = \{ I_m \mid f^R_m \geq TH^I \}
\]  

(21)

With the same idea as that in [24], to ensure that there is at least one positive image in each positive shot, the threshold \( TH^I \) is defined by

\[
TH^I = \max \{ \theta \mid \forall S_l \in S^+, \exists I_m \in S_l, f^l_m \geq \theta \} = \min \max_{S_l \in S^+} f^l_m
\]  

(22)

Similarly, when \( f^R \) is fixed, the original cost function becomes

\[
Q(f^l) = \frac{\mu_l^2}{2} \sum_{m,n=1}^{M} W_{mn} \left( \frac{f^l_m}{\sqrt{d^l_m}} - \frac{f^l_n}{\sqrt{d^l_n}} \right)^2 + \mu_l \sum_{S_l \in S^+} H_1 \left( \max_{m \in S_l} f^l_m, 1 \right) + \mu_l \sum_{l_m \in C^+} H_2 \left( f^l_m, -1 \right) + \mu_R \sum_{l_n \in C^+} H_1 \left( \max_{l_n \in R^+} f^R_n, f^l_m \right) \]  

(23)

In this case, \( C^+ \) is determined by

\[
C^+ = \{ I_m \mid \exists R_n \in S_l, f^R_n \geq TH^R \} = \{ I_m \mid \max_{R_n \in S_l} f^R_n \geq TH^R \}
\]  

(24)

where the threshold \( TH^R \) is calculated as

\[
TH^R = \max \{ \theta \mid \forall S_l \in S^+, \exists R_n \in S_l, f^R_n \geq \theta \} = \min \max_{S_l \in S^+} f^R_n
\]  

(25)

By introducing slack variables and rewriting the cost functions (20) and (23) into the form of constrained minimization like (11), the two optimization problems can also be solved by CCCP.

It should be noticed that in the process of iterative solution, we adopt two different approaches for determining the set of possible positive images. Although it cannot be guaranteed that the image sets selected by the two approaches are always the same, the method is still reasonable. The reasons can be analyzed as follows. When designing the cost term of \( Q_{FP}(F^R) \), for each possible positive image, we try to make the largest region soft label close to the image soft label. By comparing (21), (22) and (24), (25), since

\[
\max_{R_n \in S_l} f^R_n = \max_{I_m \in S_l} \max_{R_n \in I_m} f^R_n
\]  

(26)

it can be seen that if \( \max_{R_n \in S_l} f^R_n \) and \( \sum_{R_n \in S_l} f^R_n \) are with similar values, the two approaches for determining \( C^+ \) are consistent with each other.

After the iterative process has converged, the soft labels for images and regions can be calculated. To obtain the retrieval results, the final relevance score \( \phi^S_l \) for shot \( S_l \) is determined by

\[
\phi^S_l = \omega_2 \max_{l_m \in S_l} f^l_m + (1 - \omega_2) \max_{R_n \in S_l} f^R_n
\]  

(27)

where \( \omega_2 \leq 1 \) is with the same function as \( \omega_1 \) in (14), and the shots with the largest values of \( \phi^S_l \) are returned.

### 3.4 Extensions

Similarly as the case for image retrieval, our proposed video retrieval method can also be easily extended to supervised learning and can effectively deal with out-of-sample shots. Since the basic ideas are the same as those in Section 2.5, the details are omitted here.

Moreover, shot-level graph can also be involved in our proposed optimization framework. By constructing three graphs and considering the relationships between video shots, representative images, and segmented regions, both the multilayer and the multi-instance characteristics of videos can be further explored.

### 4. Experimental Results

Two data sets are used in our experiments. One is the SIVAL (Spatially Independent, Variable Area, and Lighting) image benchmark [20], the other is the TRECVid 2007 videos for the task of high-level feature extraction [15].

#### 4.1 Experiments on SIVAL

The SIVAL image set is widely used for multi-instance learning. It consists of 25 different categories, each includes 60 images. The images in one category contain the same object photographed against highly diverse backgrounds. The object may occur anywhere in the images and may be photographed at a wide-angle or close up. All the images have been segmented, and each region is represented by a 30-dimensional feature vector.

30 independent runs are conducted for all the categories in the database. In each category, 8 positive and 8 negative images are randomly selected as labeled samples, and another 34 images are used as unlabeled samples. That is to say, there are altogether 50 images involved in graph-based semi-supervised learning, and the rest 1450 images are treated as out-of-samples. In our experiments, the method of locality-constrained linear coding [28] is adopted for constructing image-level bag-of-key-points features. To compare with other methods, we also use the area under the receiver operating characteristic curve (AUC) as the performance measure.

The parameters in our proposal are set as follows. 4000 visual words are used for locality-constrained linear coding. The parameters of Gaussian functions for constructing image-level and region-level graphs are determined by the average value of all the considered distances. For calculating the combination coefficients in the designed cost function, \( \alpha, \beta, \gamma \) are set to 1, 50, and 2, respectively.

To qualitatively demonstrate the effectiveness of our proposed method, some images with calculated region soft labels are illustrated in Figure 1. In the experiments, the
shown images belong to the user-interested categories, and they are treated as out-of-samples. From the figure we can see that the regions corresponding to the user-interested objects, such as the book and the can, are assigned with larger soft labels, while the labels of the background regions are usually quite small.

Next we compare the performance of our proposal with other methods. The methods used for comparison include multi-instance learning based on region-level graph (GMIL) [27], support vector machine with evidence region identification (EC-SVM) [18], up-speed of semi-supervised multi-instance learning based on transductive support vector machine (UP-SSMIL) [33], as well as image-level semi-supervised multi-instance learning (MISSL) [19]. The average AUC values and the 95%-confidence intervals for our proposal and the other four methods are listed in Table 1. It can be seen that the overall performance of our proposal is the best. In all the 25 categories, our proposal achieves highest AUC values on 8 categories. Especially for “DataMiningBook”, “RapBook” and “LargeSpoon”, the AUC values of other methods can be improved by more than 8%. Besides our method, GMIL and MISSL also belong to the framework of graph-based learning. Since only image-level graph is adopted and the relationship between images and regions is not fully explored, MISSL performs worst of all. The main difference between GMIL and our proposal is that only region-level information is used in GMIL, while image-level bag-of-keypoints features are also introduced in our method. The superiority of the latter demonstrates the effectiveness of combining both image- and region-level information together. The basic idea of both EC-SVM and UP-SSMIL is first to map all the images into a new space, and then to conduct classification by support vector machine. Their inferior to GMIL and our proposal shows the ability of region-level graph to represent data relationship.

As far as computational load is concerned, we compare our proposal with GMIL due to their similar optimization frameworks. In GMIL, the variables to be optimized are the region soft labels, while we also need to calculate the image soft labels in our method. Generally speaking, the region number is much larger than the image number. Therefore, our proposal may cost more time than GMIL, but the difference is not significant.

4.2 Experiments on TRECVID

To demonstrate the effectiveness of our proposal for video retrieval, the data utilized in high-level feature extraction of TRECVID 2007 are adopted. There are totally 50-hour videos for development, and another 50-hour videos are used for test. 20 high-level features are evaluated in TRECVID, but some of them are not suitable for object-based retrieval. Therefore, we choose the concept of “airplane” for conducting experiments.

The positive and negative training samples are quite unbalanced for “airplane”. There are less than 60 positive images, while the number of negative images is larger than 20000. To deal with the problem, we first divide the negative image set into several parts, each of which is with the similar size as that of the positive image set. Then we conduct graph-based learning on all the positive images and one part of the negative images. At last, we involve all the calculated image and region soft labels into two ensemble graphs.

Considering the task setting of TRECVID, supervised learning is adopted, and all the test data are not used in the training process. The method of locality-constrained linear coding is also utilized to construct image-level features, and the adopted region features include color histograms in the HSV space, color moments in the LUV space, as well as coarseness and directionality vectors. All the parameters are set to the same values as those for image retrieval. For evaluation, the average precision (AP) is used as the performance measure.

First, we consider the case when training labels are assigned in image level. This is essentially a problem of image retrieval, and the only difference is that a list of shots are returned as the results. The AP of our proposal is about 0.084. While in the TRECVID evaluation results [15], the median value of all the participated runs is about 0.02, and the AP of the 10th-place run is about 0.09. From the comparison it can be seen that the performance of our proposal is promising, and we hope more satisfactory results can be
will be improved, and more appropriate parameters will be

Next, by our proposal in Section 3, we address the problem in which only shot-level training labels are available. Since it is not known which image in a positive shot is relevant to the user’s query, less information can be utilized for retrieving video shots. In this case, the performance of our method almost does not drop, and the obtained AP is also about 0.084, which demonstrates the effectiveness of our proposed optimization framework.

5. CONCLUSIONS AND FUTURE WORK

In this paper, in the framework of graph-based multi-instance learning, a novel method is proposed for object-based image retrieval. Two graphs are constructed, one is in image level, and the other is in region level. In order to consider the graph structures, the training label information, as well as the relationship between each image and its segmented regions, a suitable cost function is designed, and the image and region soft labels are calculated by solving the optimization problem. Furthermore, our proposal is extended to video retrieval. By exploring both the multi-layer structure and the multi-instance relationship in video shots, our proposal can effectively deal with the case when training labels are only assigned in shot level.

Our future work mainly includes the following two aspects. (1) The approach dealing with unbalanced training samples will be improved, and more appropriate parameters will be selected for video retrieval; (2) Shot-level features will be involved in our method to further enhance the performance.

6. REFERENCES


