Integrated Graph-based Semi-supervised Multiple/Single Instance Learning Framework for Image Annotation

Jinhui Tang [†], Haojie Li [†], Guo-Jun Qi [‡], Tat-Seng Chua [†]
[†] National University of Singapore [‡] University of Science & Technology of China
{tangjh, lihj, chuats}@comp.nus.edu.sg, qgj@mail.ustc.edu.cn

ABSTRACT

Recently, many learning methods based on multiple-instance (local) or single-instance (global) representations of images have been proposed for image annotation. Their performances on image annotation, however, are mixed as for certain concepts the single-instance representations of images are more suitable, while for some other concepts the multiple-instance representations are better. Thus in this paper, we explore an unified learning framework that combines the multiple-instance and single-instance representations for image annotation. More specifically, we propose an integrated graph-based semi-supervised learning framework to utilize these two types of representations simultaneously, and explore an effective and computationally efficient strategy to convert the multiple-instance representation into a single-instance one. Experiments conducted on the Coral image dataset show the effectiveness and efficiency of the proposed integrated framework.

Categories and Subject Descriptors: H.3.1 [Information Storage and Retrieval]: Content Analysis and Indexing – *indexing methods*.
General Terms: Algorithms, Experimentation.
Keywords: Image annotation, Multiple/single instance learning

1. INTRODUCTION

With the advent of digital imagery, the digital image collections have grown rapidly in recent years. How to index and search these images effectively and efficiently is an increasing urgent research issue in the multimedia community. Several typical search models use image samples as queries but many users found that the simple query images cannot represent their query demands. Most users prefer to search images by textual queries such as "find me images of tigers in the grass" [7], so keywords describing the images are required to rank images. Manual annotation is a direct way to obtain these key-words. However, it is labor-intensive and time-consuming. Thus automatic image annotation is necessary for efficient image search.

Many variety of learning methods have been proposed recently for automatic image annotation. While some methods employ purely supervised learning [4] or semi-supervised learning [15] with the single-instance (SI) representations of images. Most methods use multiple regions to represent each image and inference models are learned from the multiple-instance (MI) representations [8][2][3]. Which representation is most suitable for detecting the semantics in the images is an important problem. In addition, the suitable representation is also dependent on the types of concepts to be detected in the images. While many semantic concepts are more closely related to regions such as "car", "tiger" and "flowers", other concepts may relate more to the entire images such as "garden" and "beach".

For MI representations, each image is deemed as a labeled bag with multiple instances, usually comprising the segmented regions of that image. Labels (or concepts) are attached to the bags while the labels of instances are hidden. The bag label is related to the hidden labels of the instances as follows: the bag is labeled as positive if any instance in it is positive, otherwise it is labeled as negative. MI learning [6] is a type of learning algorithms to tackle the annotation problems with MI representations. Many approaches are proposed to solve the MI learning problem, and some of them are based on the well-known diverse density framework [8]. In this paper, the diverse density framework is also used to convert the MI representations of images into the SI representations.

Since labeled samples for image annotation typically come from the users during an interactive session, it is also important to obtain good results speedily using a very small amount of labeled data. Semi-supervised learning [1], which aims to learn from both labeled and unlabeled data with certain assumptions, are promising to build more accurate models than those that are achievable by using purely supervised learning methods. As a major family of semi-supervised learning, the graph-based methods have attracted more and more recent research. Many works on this topic are reported in the literature of machine learning community [19]. Some of them have been applied to image and video annotation [15][11][13].

Recently some research efforts were conducted to combine MI learning and semi-supervised learning. Rahmani *et al.* [9] proposed a MI semi-supervised learning method by transforming any MI problem into an input for a graph-based SI semi-supervised learning method that encodes the MI aspects of the problem simultaneously working at both the bag and instance levels. In [16], the authors decoupled the infer-

MM'08, October 26–31, 2008, Vancouver, British Columbia, Canada.

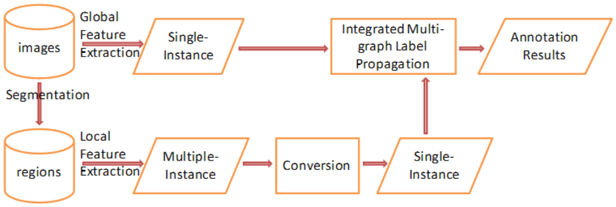

Figure 1: The integrated multiple/single instance learning framework.

ring and training stages by using random walks and SVM, and converts MI learning to a supervised learning problem. Tang et al. [12] proposed a semi-supervised MI learning method to rank natural scenes according to their typicality degrees.

To the best of our knowledge, existing learning based image annotation methods, including supervised MI learning, semi-supervised SI learning and semi-supervised MI learning, such as the aforementioned methods, just used one type of representation, and no reported methods have combined MI and SI representations in a unified framework. We believe that integrating the two types of representations will significantly improve the annotation performance. Accordingly, we propose an integrated graph-based semi-supervised multiple/single instance learning framework. The advantages of this framework for image annotation are threefold: (1) integrating the MI and SI representations for images will improve the annotation performance; (2) using a simplified diverse-density-style strategy to search the prototype for each concept, which operates only on existing instances instead of searching in the whole feature space, will result in good computational efficiency; and, (3) multi-graph based semi-supervised learning is used to integrate the multiple representations and incorporate the labeled and unlabeled data simultaneously. Experiments conducted on Coral dataset show that the proposed integrated method outperforms the normal MI and SI methods, and the conversion from MI representation to SI representation is effective and computationally efficient.

2. INTEGRATED GRAPH-BASED LABEL PROPAGATION

The proposed integrated graph-based learning framework is shown in Figure 1. First, images are segmented into regions and local features are extracted from the regions, and each image is represented with MI representation. Meanwhile, global features are extracted from the original (not segmented) images, and the features of each image form a SI representation. To integrate the MI representation and SI representation into a unified framework, we introduce a diverse-density based strategy to convert the MI representation into another SI representation. A multi-graph based label propagation method is used to integrate the two types of representations to infer the labels of the unlabeled images.

2.1 Multiple-Instance to Single-Instance Conversion

The conversion from MI representation to SI representation involves finding the prototype instance from multiple instances for each concept. It is similar to the strategy in MILES [3], and diverse-density framework is applied to find the prototype. The main difference is that we just search the maximal diverse-density point in the existed instances of positive bags but not the whole feature space, and thus the computational cost is significantly reduced.

Denote a certain positive bag as $\mathbf{B}_i^+$ and its j-th instance as $\mathbf{x}_{ij}^+$ (j=1,...,n_i^+), where n_i^+ is the number of instances in the bag B_i^+. Similarly, we use $\mathbf{B}_i^-$, $\mathbf{x}_{ij}^-$ and n_i^- respectively to represent a negative bag, its j-th instance and the number of instances in $\mathbf{B}_i^-$. In some cases we do not need to differentiate the positive and negative bags, we simply use $\mathbf{B}_i$ and n_i to denote a bag and the number of its instances. All instances are in a d-dimensional low-level feature space $\mathcal{R}^d$. We use l^+ and l^- to denote the numbers of positive and negative bags respectively. For convenience, we also use $\mathbf{x}_p$ (p=1,...,n, n=$\sum_{i=1}^{l^+} n_i^+ + \sum_{i=1}^{l^-} n_i^-$) to denote the set of all instances.

Diverse-density was proposed based on the assumption that there exists a single prototype representing each semantic concept. Other individual instances can then be annotated according to the prototype. For each concept c, the diverse-density method aims to find a point $\mathbf{x}^c$ in the feature space that maximizes the probability that the point $\mathbf{x}$ is the prototype given the training bags [8]:

$$\mathbf{x}^c = argmax_{\mathbf{x} \in \mathcal{R}^d} \prod_{i=1}^{l^+} Pr(\mathbf{x}|\mathbf{B}_i^+) \prod_{i=1}^{l^-} Pr(\mathbf{x}|\mathbf{B}_i^-), \qquad (1)$$

This strategy needs to search the whole feature space to find the prototype point for each concept. The computational cost of such searching process will be very high. To achieve the computational efficiency, we operate only on the set of likely positive instances instead of searching in the whole space, so the candidate prototype is restricted to within the instances $\mathbf{x}_p$ in the positive training bags:

$$\mathbf{x}^c = argmax_{\mathbf{x}_p \in \bigcup \mathbf{B}_i^+} \prod_{i=1}^{l^+} Pr(\mathbf{x}_p|\mathbf{B}_i^+) \prod_{i=1}^{l^-} Pr(\mathbf{x}_p|\mathbf{B}_i^-), \qquad (2)$$

where the probability $Pr(\mathbf{x}_p|\mathbf{B}_i^+)$ and $Pr(\mathbf{x}_p|\mathbf{B}_i^-)$ are estimated using the noise-or model [8]:

$$Pr(\mathbf{x}_p|\mathbf{B}_i^+) \propto 1 - \prod_j (1 - exp(-\frac{dis(\mathbf{x}_{ij}^+, \mathbf{x}_p)^2}{\sigma^2})), \qquad (3)$$

$$Pr(\mathbf{x}_p|\mathbf{B}_i^-) \propto \prod_j (1 - exp(-\frac{dis(\mathbf{x}_{ij}^-, \mathbf{x}_p)^2}{\sigma^2})), \qquad (4)$$

where σ is the scaling parameter and metric $dis(\cdot,\cdot)$ is the L_1 distance. We employ the L_1 distance instead of the widely-used L_2 distance since it has been shown that the L_1 distance can better approximate the perceptual difference of visual features [10].

For a given concept class $\mathcal{C}$=$\{c_1, c_2, ..., c_m\}$, where m is the number of give concepts, we can obtain m representation prototypes $\{\mathbf{x}^1, \mathbf{x}^2, ..., \mathbf{x}^m\}$. Using these prototypes, every bag $\mathbf{B}_i$ can be mapped into a SI representation as

$$\mathbf{b}_i = [s(\mathbf{x}^1, \mathbf{B}_i), s(\mathbf{x}^2, \mathbf{B}_i), ..., s(\mathbf{x}^m, \mathbf{B}_i)], \qquad (5)$$

where $s(\mathbf{x}^c, \mathbf{B}_i) = min_{\mathbf{x}_{ij} \in \mathbf{B}_i}\{dis(\mathbf{x}^c, \mathbf{x}_{ij})\}$, $dis(\cdot,\cdot)$ is the L_1 distance. This mapping is different from the mapping strategy in diverse-density framework since diverse-density involves scaling parameters into the mapping, which is hard to make the optimal choice.

2.2 Multi-Graph based Label Propagation

Before discussing the multi-graph based label propagation, we first introduce some basic notations: let $\mathcal{X} = \{I_1, ..., I_l, I_{l+1}, ..., I_N\}$ be a set of N image samples. For each concept, the first l image samples are labeled as $\mathbf{y}_{\mathcal{L}} = [y_1, y_2, ..., y_l]^T$ with $y_i \in \{1, 0\}$ $(1 \leqslant i \leqslant l)$ and the remaining image samples are unlabeled. The vector of the predicted labels of all samples is represented as $\mathbf{f}$, which can be split into two blocks after the l-th row: $\mathbf{f} = \left[\mathbf{f}_{\mathcal{L}}^{\mathtt{T}}, \mathbf{f}_{\mathcal{U}}^{\mathtt{T}}\right]^{\mathtt{T}}$, where $\mathtt{T}$ represents the matrix transpose. Consider a connected undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with the vertex set $\mathcal{V}$ corresponding to the N image samples. $\mathcal{V} = \mathcal{L} \cup \mathcal{U}$, where the vertex set $\mathcal{L} = \{1, ..., l\}$ contains labeled points and the vertices in set $\mathcal{U} = \{l + 1, ..., l + u\}$ are unlabeled ones. The edges $\mathcal{E}$ are weighted by the $n \times n$ pairwise similarity matrix.

Besides the multiple regions' features, we can also extract the global features from the entire image sample, which form a SI representation $\mathbf{g}_i$ of each image. So we have two types of representations for the dataset: $\{\mathbf{g}_i\}$ and $\{\mathbf{b}_i\}$, where $\mathbf{b}_i$ is the prototype representation of sample I_i as defined in Eqn. (5). Then two graphs $\mathcal{G}^{\mathbf{g}}$ and $\mathcal{G}^{\mathbf{b}}$ are constructed according to image sets, which include the global representation and the one embedded with local representation respectively. We assume that the two graphs are represented by affinity matrices $\mathbf{W}^{\mathbf{g}}$ and $\mathbf{W}^{\mathbf{b}}$ respectively, with $W_{ij}^{\mathbf{g}}$ and $W_{ij}^{\mathbf{b}}$ represent the pairwise similarity between the i-th and j-th image sample.

Then according to the theory of graph-based semi-supervised learning [19], the label inference problem becomes the problem of minimizing the following cost function:

$$C(f) = \frac{1}{2}[\alpha \sum_{i,j} \frac{W_{ij}^{\mathbf{g}}}{D_{ii}^{\mathbf{g}}}(f_i - f_j)^2 + (1 - \alpha) \sum_{i,j} \frac{W_{ij}^{\mathbf{b}}}{D_{ii}^{\mathbf{b}}}(f_i - f_j)^2] + \mu \sum_{i \in \mathcal{L}} (f_i - y_i), \quad (6)$$

where $D_{ii}^{\mathbf{g}} = \sum_j W_{ij}^{\mathbf{g}}$ and $D_{ii}^{\mathbf{b}} = \sum_j W_{ij}^{\mathbf{b}}$. The first term of the right side of (6) indicates that the labels of nearby samples should not change too much according to the structure of graph $\mathcal{G}^{\mathbf{g}}$, while the second term indicates that the labels of nearby samples should not change too much according to the structure of graph $\mathcal{G}^{\mathbf{b}}$. The third term requires that the inference function be consistent to the initial labels assignment.

Adding the constraint that the labels of annotated samples will not change in the label propagation procedure, i.e. $f_i \equiv y_i (1 \leqslant i \leqslant l)$, the optimal inference function becomes:

$$\begin{gathered} f^* = argmin_f \{\tfrac{1}{2}[\alpha \textstyle\sum_{i,j} \frac{W_{ij}^{\mathbf{g}}}{D_{ii}^{\mathbf{g}}}(f_i - f_j)^2 \\ + (1 - \alpha) \textstyle\sum_{i,j} \frac{W_{ij}^{\mathbf{b}}}{D_{ij}^{\mathbf{b}}}(f_i - f_j)^2]\} \qquad (7) \\ s.t. \quad f_i \equiv y_i (1 \leqslant i \leqslant l) \end{gathered}$$

Representing this optimization problem in the matrix form gives rise to:

$$\begin{gathered} \mathbf{f}^* = argmin_{\mathbf{f}} \{\alpha \mathbf{f}^{\mathtt{T}} \mathbf{L}^{\mathbf{g}} \mathbf{f} + (1 - \alpha) \mathbf{f}^{\mathtt{T}} \mathbf{L}^{\mathbf{b}} \mathbf{f}\} \qquad (8) \\ s.t. \quad \mathbf{f}_{\mathcal{L}} \equiv \mathbf{y}_{\mathcal{L}} \end{gathered}$$

where $\mathbf{L}^{\mathbf{g}} = \mathbf{I} - (\mathbf{D}^{\mathbf{g}})^{-1}\mathbf{W}^{\mathbf{g}}$ and $\mathbf{L}^{\mathbf{b}} = \mathbf{I} - (\mathbf{D}^{\mathbf{b}})^{-1}\mathbf{W}^{\mathbf{b}}$ are the *graph Laplacians* of $\mathbf{W}^{\mathbf{g}}$ and $\mathbf{W}^{\mathbf{b}}$ respectively; $\mathbf{D}^{\mathbf{g}}$ and $\mathbf{D}^{\mathbf{b}}$ are diagonal matrices with diagonal elements $D_{ii}^{\mathbf{g}}$ and $D_{ii}^{\mathbf{b}}$; and $\mathbf{I}$ is the identity matrix.

If we regard α as a variable and solve the optimization problem with respect to both $\mathbf{f}$ and α, the solution will be trivial since the solution is: α=1 for $\mathbf{f}^{\mathtt{T}}\mathbf{L}^{\mathbf{g}}\mathbf{f} > \mathbf{f}^{\mathtt{T}}\mathbf{L}^{\mathbf{b}}\mathbf{f}$, α=0 for $\mathbf{f}^{\mathtt{T}}\mathbf{L}^{\mathbf{g}}\mathbf{f} < \mathbf{f}^{\mathtt{T}}\mathbf{L}^{\mathbf{b}}\mathbf{f}$ and α can be any value for $\mathbf{f}^{\mathtt{T}}\mathbf{L}^{\mathbf{g}}\mathbf{f} = \mathbf{f}^{\mathtt{T}}\mathbf{L}^{\mathbf{b}}\mathbf{f}$. That is to say, only the smoothest graph is reserved. Certainly this is not the optimal solution we want. Wang *et al.* [18] proposed an EM-style iterative method to solve $\mathbf{f}$ and α, however, this process made a relaxation that change α and $(1 - \alpha)$ to α^r and $(1 - \alpha)^r$. The exponential coefficient r is sensitive to noise and is hard to choose by cross validations. Meanwhile, we only have two graphs here, as discussed in both [14] and [18], we can regard α as a parameter and determine its value by cross validations.

Let $\mathbf{L} = \alpha \mathbf{L}^{\mathbf{g}} + (1 - \alpha)\mathbf{L}^{\mathbf{b}} = \mathbf{I} - \alpha(\mathbf{D}^{\mathbf{g}})^{-1}\mathbf{W}^{\mathbf{g}} - (1 - \alpha)(\mathbf{D}^{\mathbf{b}})^{-1}\mathbf{W}^{\mathbf{b}}$, and let $\mathbf{P} = \alpha(\mathbf{D}^{\mathbf{g}})^{-1}\mathbf{W}^{\mathbf{g}} + (1-\alpha)(\mathbf{D}^{\mathbf{b}})^{-1}\mathbf{W}^{\mathbf{b}}$, then the optimization problem (8) can be transformed to

$$\begin{gathered} \mathbf{f}^* = argmin_{\mathbf{f}} \{ \quad \mathbf{f}^{\mathtt{T}}(\mathbf{I} - \mathbf{P})\mathbf{f} \quad \} \qquad (9) \\ s.t. \quad \mathbf{f}_{\mathcal{L}} \equiv \mathbf{y}_{\mathcal{L}} \end{gathered}$$

Split the matrix $\mathbf{P}$ after the l-th row and l-th column, we have:

$$\mathbf{P} = \begin{bmatrix} \mathbf{P}_{\mathcal{L}\mathcal{L}} & \mathbf{P}_{\mathcal{L}\mathcal{U}} \\ \mathbf{P}_{\mathcal{U}\mathcal{L}} & \mathbf{P}_{\mathcal{U}\mathcal{U}} \end{bmatrix}. \qquad (10)$$

Then similar to the iterative solution in [11], we can obtain the optimal label vector for unlabeled image samples as:

$$\mathbf{f}_{\mathcal{U}}^* = (\mathbf{I} - \mathbf{P}_{\mathcal{U}\mathcal{U}})^{-1}\mathbf{P}_{\mathcal{U}\mathcal{L}}\mathbf{y}_{\mathcal{L}}. \qquad (11)$$

According to Eqn.(11), each image sample will be assigned a real-valued score indicating the degree that it belongs to a specific concept.

2.3 Affinity Matrix Construction

Now we have introduced the entire framework. The remaining important issue is how to construct the affinity matrices $\mathbf{W}^{\mathbf{g}}$ and $\mathbf{W}^{\mathbf{b}}$. For simplicity, we only introduce the construction of $\mathbf{W}^{\mathbf{g}}$ here, while $\mathbf{W}^{\mathbf{b}}$ can be constructed in a similar manner. The most widely used strategy to calculate the affinity matrix for a certain concept c is as follows:

$$W_{ij}^{\mathbf{g}} = \begin{cases} exp(-\frac{dis(\mathbf{g}_i, \mathbf{g}_j)^2}{2(\sigma_c^{\mathbf{g}})^2}) & i \neq j \\ 0 & i = j \end{cases}, \qquad (12)$$

which demonstrates that there is an optimal parameter $\sigma_c^{\mathbf{g}}$ for every concept c, that is to say, using this similarity measure to predict the optimal result needs to optimize m parameters. Using cross validations to select the parameters has two problems, the first one is that the computational cost is very high, and the second is that the parameters determined by cross validations are biased to the training set. To alleviate these issues, we adopt the linear neighborhood propagation [17] to calculate the affinity matrix. As there is no parameter in the linear neighborhood propagation algorithm, thus it can tackle the aforementioned problems adequately.

3. EXPERIMENTS

To evaluate the proposed integrated multiple/single instance learning framework for image annotation, we conduct experiments on the CORAL dataset with 5,000 images. For MI representations, the images are first segmented using JSEG [5] and only the regions larger than 1/25 of original image are kept. As a result, each image usually contains less than 10 regions. A set of low-level features is extracted from each region to represent an instance, including color correlogram, color moment and wavelet texture [2]. The same features are extracted from the entire image to form the SI representation. 70 concepts are selected for experimental comparisons. The dataset is separated into two parts - the first part containing 4,500 images is used for training and the second part containing 500 images is used for test. Both

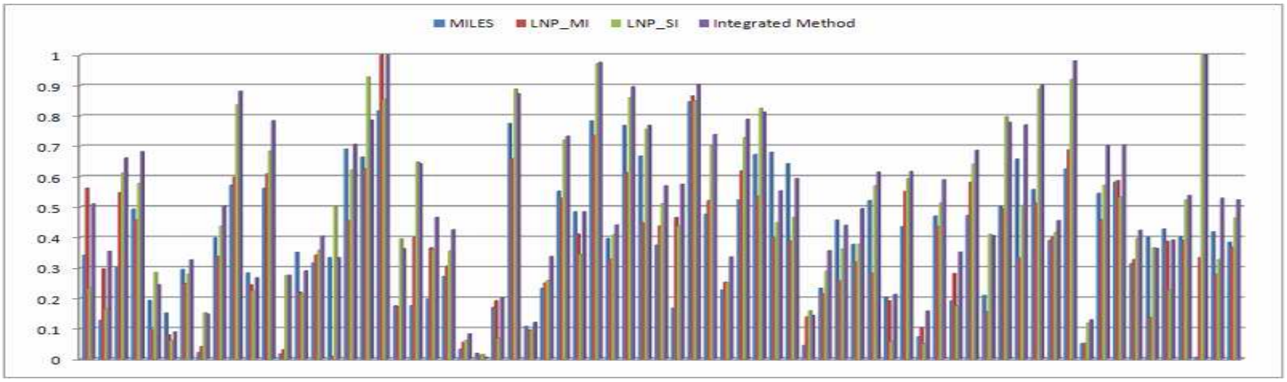

Figure 2: Experimental results over the 70 concepts.

the concepts and the data separation strategy are the same as in [2].

For each concept, the test images are ranked according to the probability that the images are relevant. The performance is measured via non-interpolated *Average Precision* (AP), a standard metric for document retrieval. We average the APs over all the 70 concepts to create the *Mean Average Precision* (MAP), which is the overall evaluation result.

We compare the results of the proposed integrated multiple/single instance learning framework with: 1) linear neighborhood propagation (LNP) [17] with only MI representation, 2) linear neighborhood propagation with only SI representation, and 3) the popular MI learning method MILES [3]. The parameter α in the framework is chosen through 5-fold cross validations. Since the feature mapping in the MILES method needs too much memory, which is hard to implement in normal PCs, we down-sample the training set to 1,000 images to run the MILES method. In the down-sampling, if the number of positive images for a certain concept is less than 600, all the positive images are reserved, otherwise the positive images will be randomly down-sampled to remain about 600 images. The negative images for all concepts are randomly down-sampled to supplement the positive images to make the number of training images be 1,000.

The experimental results are shown in Figure 2. We can see that the integrated method significantly outperforms the other three methods with MI or SI representations for most of the 70 concepts. The comparisons of MAPs and time costs for the four methods are shown in Table 1. We can see that the MAP of the integrated method is 0.524, which has improvements of 36.1%, 42.4% and 12.7% over MILES, linear neighborhood propagation with MI representation and linear neighborhood propagation with SI representation respectively. The linear neighborhood propagation with MI representation has comparable annotation performance with MILES and much smaller computational cost, so the conversion strategy from MI representation to SI representation is effective and efficient. The average time cost of the integrated method for each concept is about five minutes. It is a little slower than the linear neighborhood propagation with one type of representation, but is much faster than MILES, although the MILES is conducted when the training data is down-sampled.

4. CONCLUSIONS AND FUTURE WORK

This paper has introduced an integrated graph-based semi-supervised learning framework to utilize the MI representation and SI representation simultaneously. Experiments were conducted on the Coral image dataset to show the effectiveness and efficiency of the integrated framework for automated image annotation. However, from the experimental results we can see that the performances of methods with MI representations have not achieved the similar performance compared with the method with SI representation. Thus there needs much work to do for utilizing the MI representation. Our future work will focus on how to utilize the MI representation directly and the conversion from MI representation to SI representation.

Table 1: Comparisons of MAPs and time costs

Method	MAP	Time Cost (minutes)
MILES	0.385	20
LNP_MI	0.368	4
LNP_SI	0.465	2
Integrated method	0.524	5

5. REFERENCES

[1] O. Chapelle, A. Zien, and B. Scholkopf. *Semi-supervised Learning.* MIT Press, 2006.

[2] Y. Chen and J. Z. Wang. Image categorization by learning and reasoning with regions. *Journal of Machine Learning Research*, (5):913–939, 2004.

[3] Y. Chen, J. Bi and J. Z. Wang. MILES: Multiple-Instance Learning via Embedded Instance Selection. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 28, No. 12, 2006.

[4] C. Cusano, G. Ciocca and R. Schettini. Image Annotation Using SVM. *Proceedings of Internet Imaging*, vol. SPIE 5304, 2004.

[5] Y. Deng and B. S. Manjunath. Unsupervised segmentation of color-texture regions in images and video. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 2001.

[6] T. Dietterich, R. Lathrop, and T. Lozano-Perez. Solving the multiple instance problem with axis-parallel rectangles. *Artificial Intelligence*, 89:31–71, 1997.

[7] J. Jeon, V. Lavrenko and R. Manmatha. Automatic Image Annotation and Retrieval using Cross-Media Relevance Models. *ACM SIGIR Conference*, 2003.

[8] O. Maron and A. Ratan. Multiple-instance learning for natural scene classification. *15th International Conference on Machine Learning*, 1998.

[9] R. Rahmani and S. Goldman. Multiple-instance semi-supervised learning. *23rd International Conference on Machine Learning*, 2006.

[10] M. Stricker and M. Orengo. Similarity of color images. *Proceedings of Storage and Retrieval for Image and Video Databases (SPIE 2420)*, 2000.

[11] J. Tang, X.-S. Hua, G.-J. Qi, M. Wang, T. Mei and X. Wu. Structure-sensitive manifold ranking for video concept detection. *ACM Multimedia*, 2007.

[12] J. Tang, X.-S. Hua, G.-J. Qi and X. Wu. Typicality ranking via semi-supervised multiple-instance learning. *ACM Multimedia*, 2007.

[13] J. Tang, X.-S. Hua, G.-J. Qi, Y. Song and X. Wu. Video Annotation Based on Kernel Linear Neighborhood Propagation. *IEEE Transactions on Multimedia*, Vol.10, Issue 4, 2008.

[14] H. Tong, J. He, M. Li, C. Zhang, and W. Ma. Graph based multi-modality learning. *ACM Multimedia*, 2005.

[15] C. Wang, F. Jing, L. Zhang, and H.-J. Zhang. Image annotation refinement using random walk with restarts. *ACM Multimedia*, 2006.

[16] D. Wang, J. Li, and B. Zhang. Multiple-instance learning via random walk. *European Conference on Machine Learning*, 2006.

[17] F. Wang and C. Zhang. Label Propagation Through Linear Neighborhoods. *International Conference on Machine Learning*, 2006.

[18] M. Wang, X.-S. Hua, X. Yuan, Y. Song and L.-R. Dai. Optimizing multi-graph learning: towards a unified video annotation scheme. *ACM Multimedia*, 2007.

[19] X. Zhu. *Semi-Supervised Learning with Graphs.* PhD Thesis, CMU-LTI-05-192, 2005.