Image Annotation by Graph-Based Inference With Integrated Multiple/Single Instance Representations

Jinhui Tang, Member, IEEE, Haojie Li, Guo-Jun Qi, and Tat-Seng Chua

Abstract—In most of the learning-based image annotation approaches, images are represented using multiple-instance (local) or single-instance (global) features. Their performances, however, are mixed as for certain concepts, the single-instance representations of images are more suitable, while for others, the multiple-instance representations are better. Thus this paper explores a 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. We further explore three strategies to convert from multiple-instance representation into a single-instance one. Experiments conducted on the COREL image dataset demonstrate the effectiveness and efficiency of the proposed integrated framework and the conversion strategies.

Index Terms—Image annotation, multiple/single instance learning.

I. INTRODUCTION

ACCOMPANIED by the decreased costs for multimedia recording and storage devices, high transmission rates, and improved compression techniques, the digital image collections have grown rapidly in recent years. How to index and search for these images effectively and efficiently is an increasingly urgent research issue in the multimedia community. Several content-based search models use image samples as queries but many users found that the simple set of query images cannot represent their query demands. Most users prefer to search for images by issuing textual queries such as “find me images of tigers in the grass” [11]. To support this, keywords describing the images are required to retrieve and rank images. Manual annotation is a direct way to obtain these keywords. However, it is labor-intensive and error-prone. Thus automatic annotation of images at the semantic concept level has emerged as an important technique for efficient image search.

In recent years, many variety of learning methods have been proposed for automatic image annotation. While a few methods employ purely supervised learning [5] or semi-supervised learning [22] 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 [2, 3, 13]. Which representation is more suitable for detecting the semantics in the images is an important problem. Moreover, the suitable representation is also dependent on the types of concepts to be detected in the images. For example, while many object-oriented concepts are more closely related to regions such as “car” and “tiger”, other scene-oriented concepts may relate more to the entire images such as “garden” and “beach”.

MI representation models each image 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 of the instances in it is positive, otherwise it is labeled as negative. MI learning [7] is a type of learning algorithms to tackle the annotation problems with MI representations. Many approaches have been proposed to tackle the MI learning problem, and some of them are based on the well-known diverse-density measure [13]. In this paper, the diverse-density measure is also used in the three proposed strategies for converting the MI representations of images into SI representations.

Since labeled samples for image annotation typically come from the users during an interactive session, it is thus important to be able to obtain good results speedily using a 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 a lot of attention recently. Many works on this topic are reported in the literature of machine learning community [28] and some of them have been applied to image and video annotation [16], [18], [22].

Recently some research efforts were conducted to combine MI learning and semi-supervised learning. Rahmani et al. [14] 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 at both the bag and instance levels, but still only uses the instance-level features. In [23], the authors decoupled the inferring and training stages by using random walks and SVM, and converted MI learning to a supervised learning problem. Tang et al. [17] proposed
Fig. 1. Integrated multiple/single instance learning framework.

To the best of our knowledge, existing learning-based image annotation methods, including the 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. Since the SI and MI representations are complementary and have different strength, we believe that integrating the two types of representations will significantly improve the annotation performance.

To integrate the SI and MI representations into a unified framework, we propose a two-stage method. First, we devise efficient strategies to convert MI to SI representation. Second, we introduce a multi-graph-based label propagation method to integrate the two types of representations to infer the labels of unlabeled images. Here multi-graph-based label propagation is a semi-supervised learning strategy, which utilizes the labeled and unlabeled data simultaneously to boost the annotation performance. It has been experimentally shown to achieve better performance as compared to purely supervised learning methods.

We expect the integrated framework to offer the following three advantages: 1) integrating the MI and SI representations simultaneously for images will improve the annotation performance; 2) using the simplified diverse-density style strategies to search the prototypes for different concepts, which operate 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.

In [19], we have proposed a MI to SI conversion method that finds one prototype of instance for each concept and then maps the MI representation of every image into the space spanned by the feature vectors of the selected prototypes. In this paper, we extend the work of [19] by proposing two more effective strategies for converting the MI representation to SI representation. Experiments conducted on COREL dataset show that the proposed integrated method outperforms the normal MI and SI methods significantly, and the conversions from MI representation to SI representation are effective.

Rather than MI representation, another popular strategy for converting the local features into a global one is to cluster the local features into a global "bag-of-visual-word" representation [4], [9], [12]. However, compared to the key-points, the number of regions in an image is much smaller. If we were to convert the region features into the global bag-of-region-features representation, the resultant feature vector will be very sparse and a lot of information included in the set of region features will be lost. This will significantly degrade the annotation performance.

The rest of this paper is organized as follows. In Section II, we first present an overview of the integrated framework; while Section II-A details the three proposed conversion strategies from MI representation to SI representation, Section II-B presents the multi-graph-based label propagation to combine the multiple representations, and Section II-C describes the construction of the propagation matrix for label propagation. In Section III, we present the experimental results and discussions. The conclusion and future work are given in Section IV.

II. INTEGRATED GRAPH-BASED LABEL PROPAGATION

The proposed integrated graph-based multiple/single instance learning framework is shown in Fig. 1. First, images are segmented into regions and local features are extracted from the regions, and thus each image is represented with a MI representation. Second, meanwhile, global features are also extracted from the original (non-segmented) images, and the features of each image form a SI representation. Third, to integrate the MI and SI representations into a unified framework, we explore three diverse-density based strategies to convert the MI representation to another SI representation. Finally, a multi-graph-based label propagation method is introduced to integrate the two types of representations to infer the labels of the unlabeled images.

A. Multiple-Instance to Single-Instance Conversion

The conversion from MI representation to SI representation involves finding prototypes of the instances for the given concepts as bases to form a feature space for mapping. Three strate-
gies for the bases construction are introduced in this subsection and will be compared in the experiments. All these strategies apply diverse-density measure to find the prototypes.

We denote a certain positive bag for concept $c$ as $B_{i}^{c+}$ and its $j$th instance as $x_{j}^{c+} (j = 1, \ldots, n_{i}^{c+})$, where $n_{i}^{c+}$ is the number of instances in bag $B_{i}^{c+}$. Similarly, we use $B_{i}^{c-}$, $x_{j}^{c-}$, and $n_{i}^{c-}$, respectively, to represent a negative bag, its $j$th instance and the number of instances in $B_{i}^{c-}$. In some cases we do not need to differentiate between the positive and negative bags, we simply use $B_{i}$ 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 $F^{+}$ and $F^{-}$ to denote the numbers of positive and negative bags for concept $c$, respectively. For convenience, we also use $x_{p}$, $(p = 1, \ldots, n_{p} = \sum_{i=1}^{F^{+}} n_{i}^{c+} + \sum_{i=1}^{F^{-}} n_{i}^{c-})$ to denote the set of all instances.

**Strategy I:** The first strategy is similar to the one presented in MIILES [3]. It finds one prototype of instance for each concept and maps the MI features of every image into the space spanned by the feature vectors of the selected prototypes. 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.

Diverse-density was proposed based on the assumption that there exists an exclusive prototype of instance 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 $x^{c}$ in the feature space that maximizes the probability that the point $x$ is the prototype given the training bags [13]:

\[
x^{c} = \text{argmax}_{x \in \mathcal{R}^{d}} \prod_{i=1}^{F^{+}} P_{r}(x|B_{i}^{c+}) \prod_{i=1}^{F^{-}} P_{r}(x|B_{i}^{c-}).
\] (1)

This strategy needs to search the whole feature space to find the prototype for each concept. The computational cost of such search process will be very high. To achieve 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 $x_{p}$ in the positive training bags:

\[
x^{c} = \text{argmax}_{x_{p} \in \bigcup_{j=1}^{F^{+}} P_{r}(x_{p}|B_{i}^{c+}) \prod_{i=1}^{F^{-}} P_{r}(x_{p}|B_{i}^{c-})}
\] (2)

where the probability $P_{r}(x_{p}|B_{i}^{c+})$ and $P_{r}(x_{p}|B_{i}^{c-})$ are estimated using the noise-or model [13]:

\[
P_{r}(x_{p}|B_{i}^{c+}) \propto 1 - \prod_{j} \left(1 - \exp \left(\frac{-\text{dis}(x_{j}^{c+}, x_{p})}{\sigma^{2}}\right)\right)
\] (3)

\[
P_{r}(x_{p}|B_{i}^{c-}) \propto \prod_{j} \left(1 - \exp \left(\frac{-\text{dis}(x_{j}^{c-}, x_{p})}{\sigma^{2}}\right)\right)
\] (4)

where $\sigma$ is the scaling parameter and metric $\text{dis}(\cdot, \cdot)$ 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 [15].

For a given concept set $C = \{c_{1}, c_{2}, \ldots, c_{m}\}$, where $m$ is the number of given concepts, we obtain $m$ prototype representations $\{x^{1}, x^{2}, \ldots, x^{m}\}$. Using these prototypes, every bag $B_{k}$ can be mapped to a SI representation as

\[
b_{k} = [s(x^{1}, B_{k}), s(x^{2}, B_{k}), \ldots, s(x^{m}, B_{k})]
\] (5)

where $s(x^{c}, B_{k}) = \text{min}_{x_{ij} \in B_{k}} \{\text{dis}(x^{c}, x_{ij})\}$, $\text{dis}(\cdot, \cdot)$ is the $L_{1}$ distance. This mapping is different from the mapping strategy used in traditional diverse-density framework since the traditional one involves the use of scaling parameters in the mapping, which is hard to make the optimal choice.

**Strategy II:** In the first strategy described above we assume that each concept only has one prototype and the MI features are mapped to the same SI feature space for all concepts. However every concept may actually have more than one prototype and we should select more prototypes for each concept to form the bases of the mapped SI feature space. Thus it can select one potential prototype of instance for each positive bag, then $F^{+}$ prototypes will be obtained for each concept $c$. The selection process can be described formally as

\[
x_{k}^{c} = \text{argmax}_{x_{p} \in B_{i}^{c+}} \prod_{i=1}^{F^{+}} P_{r}(x_{p}|B_{i}^{c+}) \prod_{i=1}^{F^{-}} P_{r}(x_{p}|B_{i}^{c-})
\] (6)

Using these prototypes to form the bases of a feature space for concept $c$, the second conversion strategy maps every bag $B_{k}$ to a SI representation as

\[
b_{k}^{c} = [s(x_{1}^{c}, B_{k}), s(x_{2}^{c}, B_{k}), \ldots, s(x_{F^{+}}^{c}, B_{k})].
\] (7)

The differences between the first strategy and the second strategy for each concept $c$ can be illustrated in Fig. 2(a) and (b). The first strategy attempts to map every positive bag to a near point of the exclusive prototype for concept $c$ in the mapped feature space, while mapping every negative bag to near one of the prototypes for other concepts. The second strategy on the other hand attempts to map every positive bag to a near one of the prototypes for concept $c$ while mapping the negative bags to points far away from these prototypes.

**Strategy III:** An alternative way of constructing the bases of the new SI feature space is to use all prototypes from all concepts since all these prototypes are useful for representing images. However, this will result in the well-known problem of curse of dimensionality. Thus in Strategy III, we construct the bases for each concept $c$ using multiple prototypes for concept $c$ and a single prototype for every other concept. For concept $c$, this strategy selects the prototypes $\{x_{1}^{c}, x_{2}^{c}, \ldots, x_{F^{+}}^{c}, x^{1}, x^{2}, \ldots, x^{m_{1}}, x^{1+1}, \ldots, x^{m}\}$ to form the
bases. Then each bag \( B_i \) can be mapped to an SI representation for concept \( c \) as

\[
\mathbf{b}_i^c = \left[ s(\mathbf{x}_i^1, \mathbf{B}_i), s(\mathbf{x}_i^2, \mathbf{B}_i), \ldots, s(\mathbf{x}_i^{c-1}, \mathbf{B}_i), s(\mathbf{x}_i^c, \mathbf{B}_i), \ldots, s(\mathbf{x}_i^{m-1}, \mathbf{B}_i), s(\mathbf{x}_i^m, \mathbf{B}_i) \right]. \tag{8}
\]

The third strategy for conversion attempts to map every positive bag to near one of the prototypes for concept \( c \) in the mapped feature space while mapping every negative bag to near one of the exclusive prototypes for other concepts. The differences between this strategy and the other two strategies are illustrated in Fig. 2.

It should be noted that the second and third conversion strategies have different prototypes for each concept, so all samples are mapped to different SI feature spaces for different concepts. That is to say, we should calculate different propagation matrices for different concepts, unlike the first strategy that only needs to calculate one propagation matrix for all concepts. This makes their computational costs much higher than that of the first strategy. Thus more attention should be paid to the computational efficiency problem for converting the MI representation to SI representation for the second and third strategies.

This idea of using prototypes for feature mapping has also been employed in computer vision. In [10], the authors proposed to embed the object images into a small representative set of prototypes for object matching in different cameras. In contrast to the unknown prototypes of model images in their applications, the proposed approaches utilize the prototypes of unknown regions, where the key issue is how to find the most representative ones.

**B. Multi-Graph-Based Label Propagation**

In this subsection, we will introduce multi-graph-based label propagation which is used to integrate the two types of representations and incorporate the labeled and unlabeled data simultaneously. Before the discussion, we first introduce some basic notations. Let \( \mathcal{X} = \{I_1, \ldots, I_k, I_{k+1}, \ldots, I_N\} \) be a set of \( N \) image samples. For each concept, the first \( I \) image samples are labeled as \( \mathbf{y}_L = [y_1, y_2, \ldots, y_I]^T \) with \( y_i \in \{1, 0\} \) (\( 1 \leq i \leq I \)) 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 as: \( \mathbf{f} = \left[ \mathbf{f}_L^T, \mathbf{f}_U^T \right]^T \), where \( T \) denotes 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 \( \mathcal{L} = \{1, \ldots, I\} \) contains labeled points and the vertices in set \( \mathcal{U} = \{I+1, \ldots, I+U\} \) are unlabeled ones. The edges \( \mathcal{E} \) are weighted by the \( n \times n \) pairwise similarity matrix.

Besides the multiple regions’ features, we also extract the global features for each image, which form the SI representation \( \mathbf{g}_i \) for each image. Thus 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 (5), (7), or (8). We construct two graphs \( \mathcal{G}^g \) and \( \mathcal{G}^b \) that, respectively, correspond to the global representations and the region representations of the image sets. We assume that the two graphs are represented by their respective similarity matrices \( \mathbf{W}^g \) and \( \mathbf{W}^b \), with \( W_{ij}^g \) and \( W_{ij}^b \) represent the pairwise similarities between the \( i \)th and \( j \)th image samples.

Then according to the theory of graph-based semi-supervised learning [28], the label inference problem becomes the problem of solving the following minimization problem:

\[
f^* = \arg\min_{\mathbf{f}} \left\{ \alpha \sum_{i,j} \frac{W_{ij}^g}{D_{ii}^g} (f_i - f_j)^2 + (1 - \alpha) \sum_{i,j} \frac{W_{ij}^b}{D_{ij}^b} (f_i - f_j)^2 \right\}
\tag{9}
\]

subject to \( f_i \equiv y_i (1 \leq i \leq I) \)

where \( D_{ii}^g = \sum_j W_{ij}^g \) and \( D_{ij}^b = \sum_j W_{ij}^b \). The first term of the right side of (9) indicates that the labels of the nearby samples should not change too much according to the structure of graph \( \mathcal{G}^g \), while the second term indicates that the labels of nearby samples should not change too much according to the structure of graph \( \mathcal{G}^b \). The constraint requires that the labels of the annotated samples will not change in the label propagation procedure.
Representing this optimization problem in the matrix form gives rise to

\[
\mathbf{f}^* = \arg \min_{\mathbf{f}} \left\{ \alpha \mathbf{f}^T \mathbf{L}^b \mathbf{f} + (1 - \alpha) \mathbf{f}^T \mathbf{L}^b \mathbf{f} \right\}
\]

subject to \( \mathbf{f}^T \mathbf{y} = \mathbf{y} \),

\[
\text{subject to } \mathbf{f}^T \mathbf{y} = \mathbf{y}
\]

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

If we regard \( \alpha \) as a variable and solve the optimization problem with respect to both \( \mathbf{f} \) and \( \alpha \), the solution will be trivial since the solution is: \( \alpha = 1 \) for \( \mathbf{f}^T \mathbf{L}^b \mathbf{f} > \mathbf{f}^T \mathbf{L}^b \mathbf{f} \), \( \alpha = 0 \) for \( \mathbf{f}^T \mathbf{L}^b \mathbf{f} < \mathbf{f}^T \mathbf{L}^b \mathbf{f} \) and \( \alpha \) can be any value for \( \mathbf{f}^T \mathbf{L}^b \mathbf{f} = \mathbf{f}^T \mathbf{L}^b \mathbf{f} \).

That is to say, only the smoothest graph is reserved. Certainly this is not the optimal solution we want. Wang et al. [26] proposed an EM-style iterative method to solve \( \mathbf{f} \) and \( \alpha \). However, this process made a relaxation that changes \( \alpha \) and \( (1 - \alpha) \) to \( \alpha^r \) and \( (1 - \alpha)^r \). The exponential coefficient \( r \) is sensitive to noise and is hard to choose. Meanwhile, we only have two graphs here, as discussed in both [20] and [26], we can regard \( \alpha \) as a parameter and determine its value by cross validations.

Let \( \mathbf{P} = \alpha \mathbf{P}^b + (1 - \alpha) \mathbf{P}^b = \alpha (\mathbf{D}^b)^{-1} \mathbf{W}^b + (1 - \alpha) (\mathbf{D}^b)^{-1} \mathbf{W}^b \), then the optimization problem (10) can be transformed to

\[
\mathbf{f}^* = \arg \min_{\mathbf{f}} \left\{ \mathbf{f}^T (\mathbf{I} - \mathbf{P}) \mathbf{f} \right\}
\]

subject to \( \mathbf{f}^T \mathbf{y} = \mathbf{y} \).

Split the matrix \( \mathbf{P} \) after the \( l \)th row and \( l \)th column, we have

\[
\mathbf{P} = \begin{bmatrix}
\mathbf{P}_{L \times L} & \mathbf{P}_{L \times L} \\
\mathbf{P}_{L \times L} & \mathbf{P}_{L \times L}
\end{bmatrix}
\]

Then similar to the iterative solution in [16], we can iterate \( \mathbf{f}_{iL}^{(t+1)} = \mathbf{P}_{L \times L} \mathbf{f}_{iL}^{(t)} + \mathbf{P}_{L \times L} \mathbf{y}_L \) until convergence to obtain the optimal label vector for unlabeled images as unal

\[
\mathbf{f}_{iL}^{(t+1)} = (\mathbf{I} - \mathbf{P}_{L \times L})^{-1} \mathbf{P}_{L \times L} \mathbf{y}_L.
\]

According to (13), each image sample will be assigned a real-valued score indicating the degree that it belongs to a specific concept.

\( C. \) Construction of the Propagation Matrix

Now that we have introduced the entire framework, the remaining important issue is how to construct the propagation matrices \( \mathbf{P}^b \) and \( \mathbf{P}^b \). For simplicity, we only introduce the construction of \( \mathbf{P}^b \) here, while \( \mathbf{P}^b \) can be constructed in a similar manner. The most widely used strategy is to calculate the pairwise similarity matrix \( \mathbf{W}^b \) for a certain concept \( c \) first, then normalize this similarity matrix to obtain the propagation matrix:

\( \mathbf{P}^b = (\mathbf{D}^b)^{-1} \mathbf{W}^b \). This similarity has an optimal parameter \( \sigma^b \) for every concept \( c \), that is to say, using this similarity measure in our framework needs to estimate \( m \) optimal parameters since we have \( m \) different concepts. The estimation of parameters can be done by cross validation method. However, using cross validations to select the parameters has two problems. First the computational cost is very high, and second the parameters determined are biased to the training set. To alleviate these issues, we adopt the linear neighborhood propagation (LNP) [24] to calculate the propagation matrix. As there is no parameter in the linear neighborhood propagation algorithm, it can tackle the aforementioned problems adequately.

Here we briefly introduce the process of calculating the propagation matrix \( \mathbf{P}^b \) using LNP. Roweis and Saul [25] assumed that each sample can be optimally reconstructed by the linear combination of its neighboring samples, and the optimization objective is to minimize

\[
\varepsilon = \sum_{i} \left\| \mathbf{g}_i - \sum_{j, k \in N(g_i)} W_{ij} \mathbf{g}_j \right\|^2
\]

where \( N(g_i) \) denotes the neighboring samples of \( g_i \), and \( W_{ij} \) represents the reconstruction contribution of \( g_j \) for \( g_i \). Constraining \( \sum_{j \in N(g_i)} W_{ij} = 1 \) will result in

\[
\varepsilon_i = \left\| \mathbf{g}_i - \sum_{j \in N(g_i)} W_{ij} \mathbf{g}_j \right\|^2
\]

\[
= \left\| \sum_{j \in N(g_i)} W_{ij} (\mathbf{g}_i - \mathbf{g}_j) \right\|^2
\]

\[
= \sum_{j \in N(g_i)} W_{ij} (\mathbf{g}_i - \mathbf{g}_j)^T (\mathbf{g}_i - \mathbf{g}_j)
\]

\[
= \sum_{j, k \in N(g_i)} W_{ij} W_{ik} (\mathbf{g}_i - \mathbf{g}_j)^T (\mathbf{g}_i - \mathbf{g}_k)
\]

\[
= \sum_{j, k \in N(g_i)} W_{ij} W_{ik} (\mathbf{g}_i - \mathbf{g}_j)^T (\mathbf{g}_i - \mathbf{g}_k)
\]

where \( G_{ij} \) is the \((j, k)\)th element in the local Gram matrix \( G^b = (\mathbf{g}_i - \mathbf{g}_j)^T (\mathbf{g}_i - \mathbf{g}_k) \) of \( g_i \). Adding another constraint that every reconstruction coefficient should be nonnegative, the optimal reconstruction coefficients can thus be obtained by solving the following \( n \) standard quadratic programming problems:

\[
\min_{W_{ij}} \sum_{i, j, k} W_{ij} C_{ijk} W_{ik}
\]

subject to \( W_{ij} \geq 0 \) and \( \sum_j W_{ij} = 1 \).

\[
LNP \text{ further assumes that the sample’s label can also be reconstructed by its neighbors’ labels linearly using the same reconstruction coefficients, so we can use these construction coefficients to construct the propagation matrices } \mathbf{P}^b \text{ and } \mathbf{P}^b \text{ in our framework with}
\]

\[
P^b_{ij} = \left\{ \begin{array}{ll}
W_{ij} & j : x_j \in N(x_i) \\
0 & \text{otherwise}
\end{array} \right.
\]

(17)
and

\[ p_{ij}^b = \begin{cases} W_{ij}^b & \text{j : } x_j \in \mathcal{N}(x_i) \\ 0 & \text{else} \end{cases} \quad (18) \]

It is worth mentioning here that we have constrained \( \sum_j W_{ij}^b = 1 \) and \( \sum_i W_{ij}^b = 1 \). Thus in this propagation matrix construction strategy, we do not need to normalize \( W^b \) and \( W^b \).

For clarity, we summarize the flowchart of the whole framework in Algorithm 1.

---

**Algorithm 1: Graph-Based Label Propagation with Integrated Multiple/Single Instance Representations**

1) Extract the MI and SI features for the images to form two representation sets \( \{ \mathbf{B}_i \} \) and \( \{ \mathbf{g}_i \} \).
2) Convert the MI representation set \( \{ \mathbf{B}_i \} \) to a SI representation set \( \{ \mathbf{b}_i \} \) according to (5) [or (7), or (8)].
3) Construct the propagation matrices \( \mathbf{P}^b \) and \( \mathbf{b}^b \) as shown in (17) and (18). Then the final propagation matrix can be constructed as \( \mathbf{P} = \alpha \mathbf{P}^b + (1 - \alpha) \mathbf{P}^b \).
4) Split \( \mathbf{f} \) into two parts after the \( l \)th row, and split \( \mathbf{P} \) into four parts after the \( l \)th row and the \( l \)th column.
5) Iterate \( \mathbf{U}^{(l+1)} = \mathbf{P} \mathbf{U}^{(l)} \mathbf{Y}^{(l)} + \mathbf{P} \mathbf{U}^{(l)} \mathbf{Y}^{(l)} \) until convergence. We now obtain the optimal label vector for unlabeled image samples.

---

**III. EXPERIMENTS**

**A. Dataset**

To evaluate the proposed integrated multiple/single instance learning framework for image annotation, we conduct experiments on the COREL dataset with 5000 images [8]. For MI representations, the images are first segmented using JSEG [6] and only the regions larger than 1/25 of the original image are kept. As a result, each image usually contains less than ten regions. A set of low-level features is extracted from each region to represent an instance; the low-level features used include color moment, color correlogram 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 4500 images is used for training and the second part containing 500 images is used for testing. Both the concept set and the data separation strategy are the same as that used in [8].

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 used for document retrieval. We average the APs over all the 70 concepts to create the mean average precision (MAP).

**B. Methods for Comparison**

We compare the performances of the following five approaches to demonstrate the effectiveness of the proposed representation conversion strategies: 1) the popular MI learning method MILES [3]; 2) linear neighborhood propagation (LNP) [24] that uses bag-of-region-features (denoted as BORF); 3) LNP that uses only MI representation with the first strategy for representation conversion (referred to as MI_R_1); 4) LNP that uses only MI representation with the second strategy for representation conversion (referred to as MI_R_2); and 5) LNP that uses only MI representation with the third strategy for representation conversion (referred to as MI_R_3).

To illustrate the advantages of the integrated methods as compared to methods that uses only SI or MI representation, we compare the performances of the following three approaches: 1) LNP that uses only SI representation (denoted as SI_R); 2) MI_R_1; and 3) integrated label propagation with the first strategy for representation conversion (denoted as Integrated_1).

We also conduct the following six approaches to demonstrate the advantages of the integrated methods as compared to other methods with combined features: 1) support vector machine (SVM) [21] that uses the combined features from the SI representation and MI representation with the first strategy for representation conversion (referred to as SVM); 2) transductive support vector machine (TSVM) [21] that uses the combined feature from the SI representation and MI representation with the first strategy for representation conversion (referred to as TSVM); 3) LNP using the combined feature from the SI representation and BORF (referred to as SI+BORF); 4) Integrated_1; 5) integrated label propagation with the second strategy for representation conversion (denoted as Integrated_2); and 6) integrated label propagation with the third strategy for representation conversion (referred to as Integrated_3).

All these approaches are executed on a PC with Intel Core2 2.4 G CPU and 4 G memory.

The parameters \( \sigma \) and \( \alpha \) in the framework are chosen through 5-fold cross validations. We first use only MI representation to determine the optimal \( \sigma \) for each concept, where 4000 out of 4500 images are randomly selected from the training set for training and the rest 500 images for validation. While \( \sigma \) is selected in the set \{1, 1.5, 2, 4, 6, 8\}, respectively, with the average precision for each concept calculated on the validation set. We repeat this process five times and sum the average precision for each concept and each fixed \( \sigma \). Finally the \( \sigma \) corresponds to the largest sum of average precision is selected for every concept. Next, for each concept we fix the \( \sigma \) to the selected optimal value and use both the MI and SI representations to determine the optimal \( \alpha \). The selection process is similar to that for selecting \( \sigma \) while \( \alpha \) is selected from the set \{0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8\}. For the bag-of-region-features, we cluster the region features into \( \ell \) clusters to generate the global BORF features. In our experiments, \( \ell \) is selected from \{10, 20, 30, 50, 100\}. We choose the \( \ell \) that corresponds to the best performance for comparison. We use RBF kernel for SVM and TSVM and the parameters are chosen by 3-fold cross validations.

**C. Experiment 1: MILES versus MI_R_1**

Since the MILES method needs too much memory for implementation and is hard to implement in normal PCs, we downsample the training set to run the MILES method. During down-
sampling, if the number of positive images for a certain concept is less than 20, all the positive images are reserved. Other image samples are randomly down-sampled with the probability of 0.2. Finally we obtain a down-sampled training set with 1087 images.

The performances of MILES versus MI_R_1 using the downsampled training set are presented in Fig. 3; with the MAPs and average executing time per concept being shown in Table I. The results demonstrate that MI_R_1 has superior performance as compared to MILES in terms of both effectiveness and efficiency. Thus we can conclude that the first conversion strategy is effective and efficient.

### D. Experiment 2: Comparisons Among Other Methods That Use Only Region Features

In this experiment, we exclude MILES and compare the performances of other four approaches with region features using the entire training set as shown in Fig. 4. We can see that for almost all of the 70 concepts, the proposed three conversion strategies outperform the bag-of-region-features. Meanwhile, MI_R_2 and MI_R_3 outperform MI_R_1 remarkably for most concepts.

The MAPs and average executing time per concept for the four approaches are tabulated in Table II. MI_R_1 achieves a MAP of 0.369, which is significantly better than that obtained by BORF. As discussed before, this is because BORF approach converts the region features into a global bag-of-region-feature representation, the resulting feature vector is very sparse and a lot of information included in the set of region features is lost. This will significantly degrade the performance. We also note that MI_R_2 obtains an MAP of 0.418, which has improvement of 13.6% over MI_R_1. Furthermore, MI_R_3 achieves an MAP of 0.453, which has improvements of 23.1% and 8.37% over MI_R_1 and MI_R_2, respectively. These comparisons demonstrate that the second and third conversion strategies are successively more effective than the previous approaches.

The average executing time of MI_R_1 for each concept using the entire training set is about four minutes. It is a little slower than using BORF. One thing we should notice here is that although the methods using the second and third conversion strategies are more effective than those methods using the first conversion strategy, they require much more computational efforts. Thus we may need to employ parallelization techniques or approximation methods to accelerate the label propagation procedures. How to reduce the computational complexity for these two strategies is an important research issue to be explored in the future.

### E. Experiment 3: Comparisons Among the Integrated Method and Corresponding Methods With Single Representations

This experiment compares the performances of SI_R, MI_R, and MI_R_1 as shown in Fig. 5. We can see that...
for most of the 70 concepts, the integrated method significantly outperforms the other methods that use only the MI or SI representation. For few concepts, such as “boat” and “foals”, the performances of SI_R are better than MI_R_1. The reasons lie in two fold: 1) the SI and MI representations cannot complement each other for these concepts; and 2) the MI representation cannot represent these concepts well, thus will degrade the performance of the integrated method. Meanwhile, we can see that for most of the concepts, SI_R performs better than MI_R_1. This may be caused by the imperfect segmentations of the images.

The MAPs and average executing time for the three approaches are tabulated in Table III. The MAP of Integrated_1 is 0.524, which has improvements of 42.4% and 12.7% over...
Fig. 6. Experimental results for SVM with the integrated image representation, Integrated_1, Integrated_2, and Integrated_3, using the entire training set over the 70 concepts.

TABLE III

<table>
<thead>
<tr>
<th>Method</th>
<th>MAP</th>
<th>Average Executing Time (minutes)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SI_R</td>
<td>0.465</td>
<td>2</td>
</tr>
<tr>
<td>MI_R_1</td>
<td>0.368</td>
<td>4</td>
</tr>
<tr>
<td>Integrated_1</td>
<td>0.524</td>
<td>5</td>
</tr>
</tbody>
</table>

MI_R_1 and SI_R, respectively. This indicates that integrating the two types of representations can significantly boost the annotation performance. The average executing time for each concept of Integrated_1 using the entire training set is about 5 min. It is a little slower than linear neighborhood propagation with one type of representation, but is much faster than MILES. Thus it is both effective and reasonably efficient.

F. Experiment 4: Comparisons Among the Integrated Approaches

In this experiment, we compare the performances of the following six approaches using the integrated image representations: 1) SVM that uses the combined feature from the SI representation and MI representation with the first strategy for representation conversion (referred to as SVM); 2) TSVM that uses the combined feature from the SI representation and MI representation with the first strategy for representation conversion (referred to as TSVM); 3) LNP using the combined feature from the SI representation and BORF (referred to as SI+BORF); 4) Integrated_1; 5) Integrated_2; and 6) Integrated_3, as shown in Fig. 6. We can see that the proposed integrated approaches significantly outperforms SVM, TSVM, and SI+BORF for most of the 70 concepts, while for the rest of concepts, they have comparable performances. There is a notable outlier which is the concept “flight”, where the performances of SVM and TSVM are much better than the proposed semi-supervised learning approaches. This is because the relevant samples are few for this concept, which causes the result to be a little random. Actually there is only one relevant sample in the test set for “flight”. SVM and TSVM ranked it at the top of the result list while the other approaches ranked it at the second or third place.

The MAPs and average executing time per concept for these six integrated approaches are tabulated in Table IV. We can see that the MAP of the Integrated_1 is 0.524, which has improvements of 13.2% over SVM, 8.04% over TSVM and 7.60% over SI+BORF. The Integrated_2 and Integrated_3 again achieve further improvements of 3.25% and 4.01%, respectively, over Integrated_1 with MAPs of 0.541 and 0.545. We note that these two integrated approaches have similar performances since the SI representation can complement the information which is missing in the second feature conversion method but is included in the third conversion method. A notable outlier is the concept “harbor”. The reason for the abnormal result is the same as aforementioned: the test set only has four relevant samples for “harbor”.

Both the computational costs of SVM, TSVM, SI+BORF, and Integrated_1 are not high. Although Integrated_2 and Integrated_3 are more effective than those methods using the first
conversion strategy, they require much longer computational time. Thus how to take the parallelization techniques or approximation methods into consideration to accelerate the processes of these two strategies is an important research issue to be explored in the future.

IV. CONCLUSIONS AND FUTURE WORK

This paper introduced an integrated graph-based semi-supervised learning framework to utilize the MI and SI representations simultaneously for image annotation. To integrate the two representations into a unified framework, we explored three strategies to convert the MI representation to SI representation. Experiments conducted on the COREL image dataset demonstrated the effectiveness and efficiency of the integrated framework for automated image annotation. However, from the experimental results we can see that the computational costs of the second and third strategies for representation conversion are still high. Thus more work needs to be done to reduce the computational complexity of these two strategies. Our future work will focus on the efficiency problem of converting the MI representation to SI representation effectively.

REFERENCES


Authorized licensed use limited to: National University of Singapore. Downloaded on June 21,2010 at 17:33:58 UTC from IEEE Xplor. Restrictions apply.
Guo-Jun Qi is currently pursuing the Ph.D. degree in the Department of Electrical and Computer Engineering and Beckman Institute of the University of Illinois at Urbana-Champaign (UIUC). Before joining UIUC, he was working in Microsoft Advanced Technology Center and Microsoft Research Asia. His current research interests include multimedia information retrieval and data mining, machine learning and their applications into multimedia computing, analysis, and authoring.

Mr. Qi is a student member of the Association for Computing Machinery. He was the winner of the best paper award in the 15th ACM International Conference on Multimedia, 2007.

Tat-Seng Chua received the B.S. and Ph.D. degrees from the University of Leeds, Leeds, U.K., in 1979 and 1983, respectively.

He then joined the National University of Singapore, where he is now a full Professor at the School of Computing. He was the Acting and Founding Dean of the School of Computing during 1998–2000. He spent three years as a research staff member at the Institute of Systems Science (now I²R) in the late 1980s. His main research interest is in multimedia information processing, in particular on the extraction, retrieval, and question-answering (QA) of text and video information. He focuses on the use of relations between entities and external information and knowledge sources to enhance information processing. He is currently working on several multi-million-dollar projects: interactive media search, local contextual search, and event detection in movies. His group participates regularly in TREC-QA and TRECVID news video retrieval evaluations.

Dr. Chua is active in the international research community. He has organized and served as program committee member of numerous international conferences in the areas of computer graphics, multimedia, and text processing. He is the conference co-chair of ACM Multimedia 2005, CIVR (Conference on Image and Video Retrieval) 2005, and ACM SIGIR 2008. He serves in the editorial boards of: ACM Transactions of Information Systems (ACM), The Visual Computer (Springer), and Multimedia Tools and Applications (Springer). He is a member of the steering committee of CIVR, Computer Graphics International, and Multimedia Modeling conference series; and as member of International Review Panels of two large-scale research projects in Europe.