Multi-Label Learning With Fused Multimodal Bi-Relational Graph

Jiejun Xu, Vignesh Jagadeesh, and B. S. Manjunath, *Fellow, IEEE*

***Abstract*—The problem of multi-label image classification using multiple feature modalities is considered in this work. Given a collection of images with partial labels, we first model the association between different feature modalities and the images labels. These associations are then propagated with a graph diffusion kernel to classify the unlabeled images. Towards this objective, a novel *Fused Multimodal Bi-relational Graph* representation is proposed, with multiple graphs corresponding to different feature modalities, and one graph corresponding to the image labels. Such a representation allows for effective exploitation of both feature complementariness and label correlation. This contrasts with previous work where these two factors are considered in isolation. Furthermore, we provide a solution to learn the weight for each image graph by estimating the discriminative power of the corresponding feature modality. Experimental results with our proposed method on two standard multi-label image datasets are very promising.**

***Index Terms*—Graph-based semi-supervised learning, multi-label classification, multimodal.**

I. Introduction

IMAGE classification has a broad range of applications, including search and retrieval in large image databases. Some of the early work concerned labeling a given image with a single class label. More recently, with emergence of social media sites and sharing and annotation of multimedia files online, there is a growing interest in associating multiple labels to a single image or parts of an image. Fig. 4 illustrates the idea of multi-label image classification. For example, the image on the lower left can be classified as "lake", "sky", and "water", where each of these terms represents a different semantic concept. This motivates our interest in the following problem: *given a partially labeled database of images (from a label set $\mathcal{L}$), classify every unlabeled image i in the database with a label subset $l_i \subseteq \mathcal{L}$* (see Fig. 1).

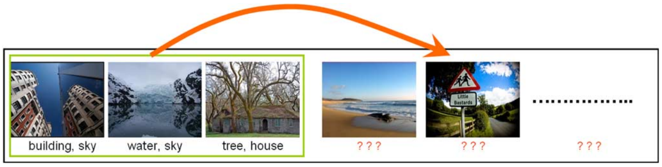


Fig. 1. Multi-label Classification: Given a partially labeled database of images, the goal is to propagate these labels and classify every unlabeled image in the database.

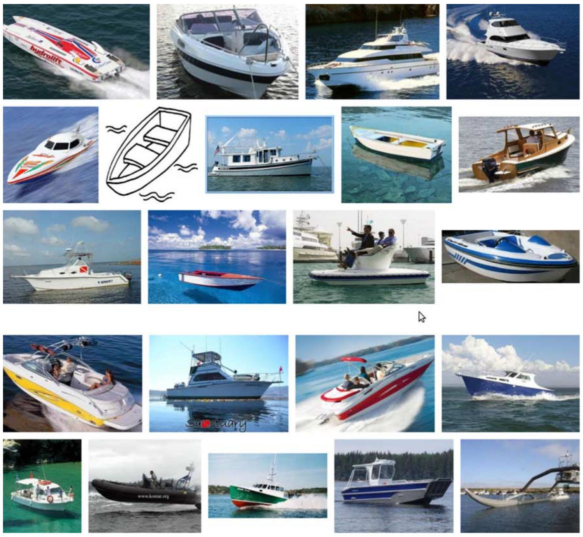


Fig. 2. Images returned with the query term "boat", showing a strong correlation with between "boat" and "ocean".

A straight forward approach to this multi-label classification problem is to convert it to a simple binary classification problem. That is to train a set of binary classifiers, one for each class, and the final labels[1] for each image is determined by combining results from all the classifiers. However, as pointed out by [1], [2], this approach treats different concepts in isolation and completely ignores the underlying correlations among them. For instance, the labels "boat" and "ocean" are likely to co-occur more frequently than the labels "boat" and "road". This can be easily observed from the top ranked image results from Google with the query term "boat", as shown in Fig. 2. To take into account dependencies among concepts, a simple extension is to enumerate all combination of the concepts, and train a binary classifier for each. An obvious drawback for such method is that, the number of combined classes increases exponentially, and the problem will be intractable if the original number of classes is large. In addition, there maybe be only very few instances in the combined classes.

Manuscript received October 09, 2012; revised April 28, 2013 and July 20, 2013; accepted August 22, 2013. Date of publication January 09, 2014; date of current version January 15, 2014. This work was supported by the following awards: NSF III 0808772, ONR #N00014-10-1-0478, and #N00014-12-1-0503. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Selcuk Candan.

J. Xu is with the Department of Computer Science, University of California at Santa Barbara, Santa Barbara, CA UC Santa Barbara, Santa Barbara, CA 93106 USA.

V. Jagadeesh and B. S. Manjunath are with the Department of Electrical and Computer Engineering, University of California at Santa Barbara, Santa Barbara, CA 93106 USA.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2013.2291218

[1] A "label" is just the name of a "class".

A common problem with typical classification problem is the small number of labeled samples available for training. Thus, as pointed out in [3], a large number of work on *semi-supervised learning* have been proposed to capitalize on the abundance of unlabeled data to improve classification performance. A recent book [4] provides an extensive survey on this subject. One of the earlier techniques developed for semi-supervised learning is Transductive SVMs (also known as $\mathrm{S}^3\mathrm{VM}$) [5]. The goal is to find a labeling of the unlabeled data, so that a linear boundary has the maximum margin on both the original labeled data and the unlabeled data. In semi-supervised learning using generative models [6], class labels of unlabeled data are treated as missing variables, and the class conditional models over the features are iteratively estimated using EM algorithms. Among other families of techniques, graph-based semi-supervised learning [7], [8] has gained significant interest due to its effectiveness and easy adaptation to various applications. Such methods define a graph where the nodes are labeled and unlabeled samples, and edges reflect the similarities. Some of the earlier works to apply graph-based semi-supervised learning on images are seen in [9]–[11].

Another important aspect of image classification is the effective use of different feature modalities. In the past, many methods have been proposed to fuse multimodal[2] features together for better visual analysis. For example, in [12], a discriminatively trained nearest neighbor model is proposed to integrate a collection of image metrics for annotation. In [13], text and visual features are fused together for ranking videos. In [14] user tags and visual feature are combined in a pairwise constrained propagation framework for clustering on web images. In [15] a model based on Conditional Random Fields (CRF) with embedded Kernelized Logistic Regression is proposed for image annotation to capture the tight interaction between different visual features and semantic context. Generally speaking, feature fusion methods are characterized into two categories, early fusion and late fusion [16]. Early fusion means representing the image data in a multimodal feature space. The simplest early fusion method consists of concatenation of features descriptors. Late fusion, on the other hand, happens at the decision level. Individual features are used in visual computation, and their results are then combined together with some kinds of aggregation function. A new type of fusion method, which utilizes graphs as a medium to combine different features, has gained significant attention recently. Promising results with this method have been reported for image clustering [14] and annotation [17]–[19]. This method is closely related to our work, and more details can be found in Section III-B.

Taking into consideration all three aspects (*label correlation, small training sample, and multimodal features*) mentioned above, we propose a novel *Fused Multimodal Bi-relational Graph* representation for multi-label image classification, with multiple graphs corresponding to different modalities in the input data space, and one graph for the labels in the output space. By doing so, feature complementariness and concept correlation are effectively and simultaneously exploited. An illustration of our proposed scheme is shown in Fig. 3.

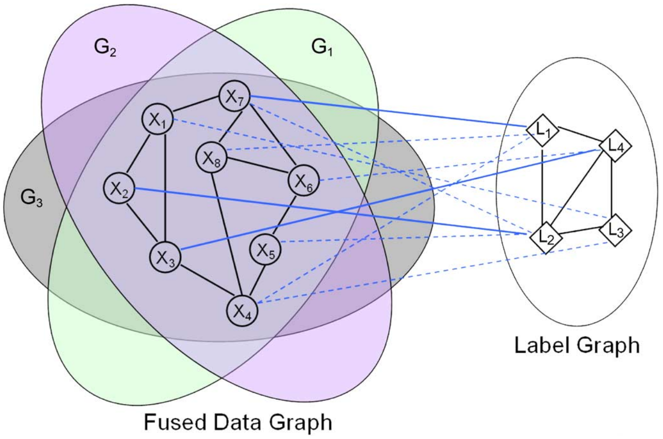


Fig. 3. An illustration of our proposed scheme. Solid black lines indicate affinity relationships among data nodes (i.e. images) X_i and label nodes L_i within their corresponding graphs. The solid blue lines across two graphs denote the initial label assignments, and the dotted lines denote the label assignments to be estimated.

The rest of this paper is organized as follows. Section II reviews related research. Section III-A presents the basis of single graph learning. Section III-B explains multimodal graph learning for combining different feature types. Section III-C presents the details of our novel *Fused Bi-relational Graph* representation for image classification. Section IV presents experimental results on standard datasets, and Section V concludes the paper.

II. Related Work

Several multi-label classification methods have been proposed to model concept correlations. A RankSVM model is proposed in [20] based on a novel definition of loss and margin for multi-label problems. A fusion-based discriminative methods is proposed in [21] to exploit correlation among classes using a general kernel function for combining text and class membership features. In [22], Multi-concept Discriminative Random Field (MDRF) is proposed to build a probabilistic model on video semantic concept detection by incorporating related concepts. In [23], a multi-label learning method based on constrained non-negative matrix factorization is presented. Their method was shown to be effective even when the number of classes is large and the size of training data is small. In [24], Correlation Label Propagation (CLP) is developed to explicitly capture the interactions between labels. Rather than treating labels independently, CLP simultaneously co-propagates multiple labels from training examples to testing examples. An algorithm based on submodular functions is also provided by the authors to solve the problem efficiently. In [25], a unified Correlated Multi-Label (CML) support vector machine is proposed to simultaneously classify labels and model their correlations in a new feature space which encode both labels and their interactions together. In [26], a transductive multi-label classification approach based on the discrete hidden Markov Random Field (MRF) is proposed to capture the interdependence among labels. Another related work based on MRF is proposed in [27] for semantic context modeling. It places special focus on parameter learning and exploring the learning power of generative models. In [28], a correlated

[2] The term “multimodal” is often used to denote different types of sensor data. In this work, we use the term in a broader sense to denote different visual descriptions computed from the image data, such as local and global features, as well as associated metadata such as tags.

linear neighborhood propagation method (CLNP) is proposed for video annotation. A special emphasis of this work is to remove false label correlations by utilizing mutual information and a manually established binary mask table. Besides these approaches, a variety of mechanisms are also used to take into account label correlations, such as maximizing label entropy [2], discriminant subspace [29], [30], directed graph [31], [32], and many others [33]–[36]

Feature fusion is also related to the area of multiple kernel learning (MKL) [37]. MKL has emerged as an effective method for combining different types of features recently. The basic idea is to learn an optimal linear combination of kernels, each of which captures a different feature representation. In [38], MKL is used to learn a combination of exponential χ^2 kernels for object detection. Besides showing the gain on detection accuracy, the authors also demonstrated that MKL is able to improve the efficiency of inference by determining a sparse selection of features. A similar MKL framework is used in [39] for image classification. Both image content and the associated tags are combined by MKL to form a stronger classifier. Their work is slightly different from others work in the sense that their training data has additional features that are absent from the test data. In [40], a hierarchical kernel is proposed to combine visual features with depth information for object recognition. In their follow-up work [41], a compact kernel descriptor is proposed to integrate both rgb color and depth information in a unified way.

III. Multi-Label Classification in a Unified Graph Framework

The proposed graph-based multi-label image classification technique represents a transductive semi-supervised learning process that diffuses the label information from a small subset of images to the rest in the graph. Similarity measures from multiple modalities are combined in a principled way to improve the classification performance. In particular, through careful construction of a bi-relational graph, label correlation constraint is exploited jointly in the diffusion process.

A. Single Graph Learning

The basic idea is to construct a graph where vertices represent images and edge weights represent similarity between two images. Image similarity can be computed using any feature distance metric, such as the Euclidean distance between two visual descriptors, or the cosine distance between two tag vectors.

In a typical image classification process, a small set of labeled training images are given first. Then classifiers are learned based on the training images, and subsequently used to predict labels of new images. In a graph setting, this is done by first assigning labels on a small set of nodes (i.e., images) in the graph, and then propagating these labels to the rest of the unlabeled nodes based on the graph structure. The keys to semi-supervised learning are the two prior assumptions of consistency: First, nearby data points are likely to share the same label; Second, data points on the same structure (cluster/manifold) are likely share the same label. Note that the first assumption is local, whereas the second one is global. Different graph-based methods have been proposed for label propagation. Generally they are formulated in a regularization framework $F^* = argmin_F\{\Omega_{smooth}(F) + \Omega_{fit}(F)\}$, where F is the to-be-learned vector containing the soft labels of the graph nodes. The first term is a loss function corresponding to the smoothness constraint on the neighboring labels. This is to say nodes which can be connected via a path through high density regions on the graph should share similar labels. The second term is a regularizer for the fitting constraint, which means that initial assigned labels should be changed as little as possible. It should be clear that the two terms are in accordance with the two prior assumptions of consistency. Several well-known methods following the same framework are: Gaussian random field and harmonic function [7], Riemannian Manifolds [42], and the Local and Global Consistency approach [8]. In this work, we follow the regularization framework by [8] due to its effectiveness in image classification.

Given a set of data points $\mathcal{X} = x_1, x_2, \ldots, x_N$, and their affinity matrix W, the objective function to classify each data with labels $\mathcal{L} = l_1, l_2, \ldots, l_K$ is defined as follows:

$$\Omega(F) = \underbrace{\frac{1}{2}\sum_{i,j=1}^{N} W_{i,j} \left\| \frac{F_i}{\sqrt{D_{ii}}} - \frac{F_j}{\sqrt{D_{jj}}} \right\|}_{Loss function} + \mu \underbrace{\sum_{i}^{N} \|F_i - Y_i\|}_{Regularizer}, \tag{1}$$

where Y is the $N \times K$ initial labeling matrix with $Y_{ij} = 1$ if x_i is labeled with l_j, $Y_{ij} = -1$ if x_i is not labeled with l_j, and $Y_{ij} = 0$ if the label of x_i is unknown; D is the diagonal normalizing matrix given by $D_{ii} = \sum_j W_{ij}$; F is the $N \times K$ target label matrix. F_{ij} is essentially a confidence value of assigning label l_j to x_i. The closed-form solution for the minimization is found to be

$$F = \left(I + \frac{1}{\mu}L\right)^{-1} Y, \tag{2}$$

where L is the *normalized Laplacian* given by $L = D^{-(1/2)}(D - W)D^{-(1/2)} = I - D^{-(1/2)}WD^{-(1/2)}$. For computational efficiency, [8] also provided an iterative solution for F along with its proof of convergence.

B. Multimodal Graph Learning

Multimodal graph learning has been used in situations in which more than one type of features can be used to measure the affinity between vertices. This is especially common in the image classification context, as the similarity between two images can be measured by different features, such as global and local visual feature and user tags. Combination of multiple features can be done through a weighted union of graphs generated by different features. Each graph resembles a weak classifier generated from a single cue, and together they form a stronger classifier.

The problem of multimodal graph learning not only needs to address label assignments, but also needs to address learning the weights α_g, which are the weighting terms used when combining the individual graphs. In a simple case, $[\alpha_1, \ldots, \alpha_G]$ can be set to equal values. But a better alternative is to learn the values of α_g by estimating the quality of each graph. Typically, the quality of a graph is characterized by the degree of smoothness. This is because the smoother a graph, the more consistent labels are assigned to nodes with respect to its intrinsic structure.

The regularization framework in (1) was extended to handle multimodal features in [17], [43] by adding a weighting coefficient to each graph. The extended loss function is defined as follows,

$$\frac{1}{2}\sum_{g=1}^{G}\alpha_g^r\left(\sum_{i,j}^{N}W_{g,ij}\left\|\frac{f_i}{\sqrt{D_{g,ii}}}-\frac{f_j}{\sqrt{D_{g,jj}}}\right\|^2\right). \qquad (3)$$

Instead of α_g, the weight coefficient is actually relaxed to α_g^r, where $r > 1$. This is to prevent the solution from reducing to the trivial case which keeps the graph from only one modality (i.e. the one with the highest smoothness degree) when $r = 1$. As $r \rightarrow 1$, better ("smoother") graph will be emphasized with higher weight. As $r \rightarrow \infty$, equal weights are assigned to all graphs. One thing to note is that, the work in [43] only focus on binary (single label) classification for video annotation, i.e. f_i is a single value as oppose to a vector.

In the next section, we augment the multimodal graph learning technique to work with multi-label classification. And more importantly, we introduce a new regularizer in the framework to emphasize label correlation constraint.

C. Multimodal Bi-Relational Graph Learning

Traditional graph-based learning techniques usually only construct data graphs based on image features. However, in order to consider both feature complementariness and label correlation at the same time, we define a *Fused Multimodal Bi-directional Graph (FMBG)* representation, which contains two main graphs corresponding to the two spaces. An illustration of our scheme is shown in Fig. 3. On the one hand, we derive a fused data graph by optimally combining individual graphs from different modalities (e.g. global and local visual feature and user tags etc.). These graphs share the same vertex set, the difference lies on their connecting edges. By fusing data graphs together, we aim to capture the inter-connection among vertices in a better way. On the other hand, we construct a second graph to capture the class correlation in the label space. Given initial label assignments for a small set of data nodes (i.e. initial connection between image nodes and label nodes), the problem is to estimate the affinities between the two types of nodes in the entire graph. The construction of the Bi-relational graph is detailed as follows.

Fused Data Graph: Following traditional graph-based techniques, the whole dataset is modeled as a graph such that the nodes correspond to data points (e.g. images), and edges correspond to the similarities between connected data points. Instead of a fully connected graph, each data point is only connected to its top-k nearest neighbors to make graph sparse for computational efficiency. Empirical study shows that a relatively small number of k is sufficient to make the graph connected, i.e. a path exists between any two vertices. To integrate multiple modalities, several graphs are constructed using different features. In our proposed work, the optimal weights to combine all the graphs together will be learnt, as shown in later section.

Label Graph: Vertices in this graph correspond to labels, and edges indicate the correlation among them. In order to construct this graph, we first define the correlation metric between labels. This can be done through a set of training instances annotated for the label set. A binary vector, whose elements correspond to training instances, is built for each label. The value of each vector element is set to either one or zero depending on whether the corresponding training instance belongs to the concept. Then we can use standard cosine similarity to indicate the correlation between two labels. Similar to the data graph, each vertex is only connected to its k nearest neighbors.

Existing graph-based semi-supervised learning frameworks attempt to minimize a cost function which takes into account two properties: smoothness of the data graph and the deviation of initial assignments. Here we introduce a third property into the regularization framework, smoothness in the label graph. Let W_g be a $N \times N$ affinity matrix denoting the data graph constructed from feature modality g with N data points, and W' be a $K \times K$ affinity matrix denoting the label graph constructed for K concepts. Let $F = (F_1, \ldots, F_N)^T = (C_1, \ldots, C_K)$ be a $N \times K$ matrix denoting the final affinity values between every image label pairs. $(C_1, \ldots, C_K)$ are the columns of F, corresponding to the K labels. Similarly let $Y = (Y_1, \ldots, Y_N)^T$ be an $N \times K$ matrix denoting the initial label assignments. There are three possible values $\{1, 0, -1\}$ for each Y_{ij}: 1 if image i is classified with label j, -1 if it does not, 0 if it is unlabeled.

As mentioned in previous section, the weight for each data graph is incorporated into our learning framework. This is inspired by the work from [19], [43], which utilize a learning method for video annotation with a single label. We extend that by introducing a new label graph constraint, and adapt it to handle multi-label image classification. This leads to the following regularization framework regarding F and α:

$$\Omega(F,\alpha) = \underbrace{\frac{1}{2}\sum_{g=1}^{G}\alpha_g^r\left(\sum_{i,j}^{N}W_{g,ij}\left\|\frac{F_i}{\sqrt{D_{g,ii}}}-\frac{F_j}{\sqrt{D_{g,jj}}}\right\|^2\right)}_{Smoothness\ constraint\ on\ the\ fused\ data\ graph} + \eta\ \underbrace{\frac{1}{2}\sum_{i,j}^{K}W'_{ij}\left\|\frac{C_i}{\sqrt{D'_{ii}}}-\frac{C_j}{\sqrt{D'_{jj}}}\right\|^2}_{Smoothness\ constraint\ on\ the\ label\ graph} + \mu\underbrace{\sum_{i}^{N}\|F_i - Y_i\|^2}_{Label\ deviation}, \qquad (4)$$

where α_g is the weight for graph g; D and D' are both diagonal matrix whose (i, i) entries equal to the sum of the i-th row of W and W', i.e. $D_{ii} = \sum_{j=1}^{N} W_{ij}$ and $D'_{ii} = \sum_{j=1}^{N} W'_{ij}$. The solution of F and α can be found by minimizing the above cost function, with the constraint that $\sum_{g=1}^{G} \alpha_g = 1$.

The first term of the (4) is the smoothness constraint on the data graph. Minimizing it means neighboring vertices should share similar labels. For instance, if two images are close to each other based on different similarity measures, they will probably have common label(s). The second term of (4) is the smoothness constraint on the label graph. Minimizing it means neighboring vertices should include similar images. For instance, if two classes are highly correlated with each other, then they are likely to co-occur in the same image, thus overall they should share a similar set of images.

η and μ are two constants controlling the trade-off of the regularization terms. If η is set to zero, it simply means that we will ignore the correlation among labels, and the formulation is reduced to semi-supervised learning on the fused data graph. The first term of (4) can be rewritten as:

$$\begin{aligned}
&\frac{1}{2}\sum_{g=1}^{G}\alpha_g^r\left(\sum_{i,j}^{N}W_{g,ij}\left\|\frac{F_i}{\sqrt{D_{g,ii}}}-\frac{F_j}{\sqrt{D_{g,jj}}}\right\|^2\right)\\
&=\frac{1}{2}\sum_{g=1}^{G}\alpha_g^r\left(\sum_{i=1}^{N}F_i^TF_i+\sum_{j=1}^{N}F_j^TF_j-2\sum_{i,j=1}^{N}\frac{W_{g,ij}F_i^TF_j}{\sqrt{D_iD_j}}\right)\\
&=\sum_{g=1}^{G}\alpha_g^r\left(\sum_{i=1}^{N}F_i^TF_i-\sum_{i,j=1}^{N}W_{g,ij}\frac{F_i^TF_j}{\sqrt{D_iD_j}}\right)\\
&=\sum_{g=1}^{G}\alpha_g^r\left(\mathrm{tr}\left(F^T\left(I-D^{-\frac{1}{2}}W_gD^{-\frac{1}{2}}\right)F\right)\right). \qquad (5)
\end{aligned}$$

The second term of the cost function can be derived into matrix form with similar steps. Thus the whole cost function in (4) can be written in a more concise form as

$$\Omega(F,\alpha)=\sum_{g=1}^{G}\alpha_g^r\left(\mathrm{tr}(F^TL_gF)\right)+\eta\mathrm{tr}(FL_cF^T)+\mu\|F-Y\|^2, \qquad (6)$$

where $L_g = I - D^{-(1/2)}W_gD^{-(1/2)}$ and $L_c = I - D'^{-(1/2)}W'D'^{-(1/2)}$. They are the *Normalized Laplacian* of data graph and label graph respectively, and both are symmetric matrices.

By applying the following matrix properties:

$$\frac{\partial\mathrm{tr}(X^TAX)}{\partial X}=(A+A^T)X,\ \frac{\partial\mathrm{tr}(\mathrm{XAX^T})}{\partial X}=X(A+A^T),$$

we can differentiate (6) with respect to both F and α as follows:

$$\begin{aligned}
\frac{\partial\Omega(F,\alpha)}{\partial F}&=\sum_{g=1}^{G}\alpha_g^r(L_gF)+\eta FL_c+\mu(F-Y)\\
\frac{\partial\Omega(F,\alpha)}{\partial\alpha_g}&=r\alpha_g^{r-1}\left(\mathrm{tr}(F^TL_gF)\right). \qquad (7)
\end{aligned}$$

The solution for F and α can be obtained using an EM style iterative process, i.e. first fix α, then solve for F, and vice versa. Starting by initialing α vector with random values which sum up to one, in each of the iterative step, F can be solved by requiring $\partial\Omega(F,\alpha)/\partial F$ to zero. With some simple algebraic steps, we have

$$\left(\sum_{g=1}^{G}\alpha_g^rL_g+\mu I\right)F+\eta FL_c=\mu Y, \qquad (8)$$

which is essentially a matrix equation with the form of $AX+XB=C$, where $A\in\mathbb{R}^{m\times m}$, $B\in\mathbb{R}^{n\times n}$, and $X,C\in\mathbb{R}^{m\times n}$. Numerical solution to the equation can be obtained from several existed software libraries, such as LAPACK [44] and the *Lyapunov* function in Matlab.

Similarly we can solve for α by fixing F with value obtained from the previous step, and require that $\partial\Omega(F,\alpha)/\partial\alpha_g = 0$. Notice that $\sum_{g=1}^{G}\alpha_g=1$, thus it is essentially a constraint minimization problem. This can be solved with traditional *Lagrange multipliers*, which gives the following:

$$\alpha_g=\frac{\left(\frac{1}{\mathrm{tr}(F^TL_gF)}\right)^{\frac{1}{(r-1)}}}{\sum_{g=1}^{G}\left(\frac{1}{\mathrm{tr}(F^TL_gF)}\right)^{\frac{1}{(r-1)}}}. \qquad (9)$$

bus, car
aeroplane, person
car, boat, person
lake, sky, water
plant_life, structures, tree, sky
clouds, sky, structures, sunset

Fig. 4. Sample images from the PASCAL VOC'07 (top row) and the MIR Flickr (bottom row) datasets along with their concept labels.

The iterative process will eventually converge due to the fact that $\Omega(F,\alpha)$ is convex with respect to both F and α. Once F is found, we can assign labels to images using simple thresholds. Basically an image with a higher value can be assigned to the corresponding class with higher confidence.

IV. Experiments

We now evaluate the effectiveness of the proposed *Fused Multimodal Bi-directional Graph (FMBG)* for multi-label image classification. Given a partially labeled set of images, the task is to classify all the unlabeled images in the set with one or more labels.

A. Experimental Settings

In the experiments, we use two multi-label datasets, PASCAL VOC'07 [45] and MIR Flickr [46]. Example images from the two datasets are shown in Fig. 4. The VOC'07 dataset contains about 10000 images, and each image is annotated with labels from a set of 20 concepts. The MIR Flickr dataset contains 25000 images, and images are labeled with two different sets of annotations, one with 24 classes (MIR24), and the other with 38 classes (MIR38). In our data graph construction, we utilize the following features[3] and distance metrics:

DenseSift Signature: Local Sift [47] descriptors are extracted densely on an image, and then quantized into a visual codebook. In addition, each image is horizontally decomposed into 3 × 1 blocks to capture the relative arrangement of features. Histogram representation of the codebook from each block is concatenated together to form an image signature. The final signature for each image has a dimension of 3000. In addition, we

[3] Pre-extracted visual and tag features for the two datasets are publicly available at http://lear.inrialpes.fr/data.

use the *Chi-square* metric to compute the distance between two image histograms as follows:

$$d(x_1, x_2) = \frac{1}{2}\sum_{k=1}^{K} \frac{(x_{1k} - x_{2k})^2}{x_{1k} + x_{2k}}. \qquad (10)$$

This metric tests the "goodness of fit" of histogram x_1 and histogram x_2, and is symmetrical. The distance differs from the commonly used sum of squared distance (SSD) by the denominator term which regularizes the effect of bins with large counts.

GIST: The GIST descriptor describes the spatial layout of an image using global features derived from the spatial envelope of an image. It encodes the texture information of horizontal or vertical lines in an image to help matching scenes with similar layouts. The Gist feature is computed on a gray scale image by convolving it with a Gabor [48] filter at different orientations and scales. This way the high and low frequency repetitive gradient directions of an image can be measured. The pixel responding scores from the filter convolutions are stored in an array, which is the GIST feature descriptor for that image. In this work, the Gist descriptors is computed using a Gabor filter at 8 orientations and 4 different scales. The results are then averaged on a 4-by-4 grid. This gives us the final descriptor a dimension of 512. The distance between two images are computed using standard Euclidean (L2) metric.

Tags: User annotations/tags associated with images are explored as an additional feature to capture the image semantics. Due to the fact that tags are generally noisy and unlimited, all the tags that appear less than a certain threshold (i.e. occur less than 8 times) are discarded. This leaves a vocabulary of 804 tags for the VOC'07 dataset and 457 tags for the MIR Flickr dataset. A binary vector is then used to represent the absence or present of each tag from the fixed dictionary. We compute the difference between two binary vectors of image x_1 and x_2 using the Cosine distance

$$d(x_1, x_2) = 1 - \frac{\sum_{t=1}^{T} x_{1t}x_{2t}}{\sqrt{\sum_{t=1}^{T}(x_{1t})^2}\sqrt{\sum_{t=1}^{T}(x_{2t})^2}}, \qquad (11)$$

which is commonly used in vector space model for text retrieval. Essentially it measures the angle between two tag vectors x_1 and x_2.

The dynamic range of the distances computed in each feature modality can be very different, and we use a heat kernel to convert distance to similarity in the range of [0,1]. In other words, the similarity between two images x_i and x_j from modality g is estimated by

$$W_{g,ij} = \begin{cases} \exp\left(-\frac{d_g(x_i,x_j)}{\sigma_g}\right), & if\ i \neq j \\ 0, & \text{otherwise} \end{cases} \qquad (12)$$

where d_g and σ_g is the distance matrix and its mean for feature modality g.

In our label graph construction, we utilize the training portion of each dataset to compute the affinity matrix for the labels. Standard cosine similarity is used to estimate the affinity between each pair of labels. Fig. 5 shows the visualization of the label correlation matrix for MIR(38).

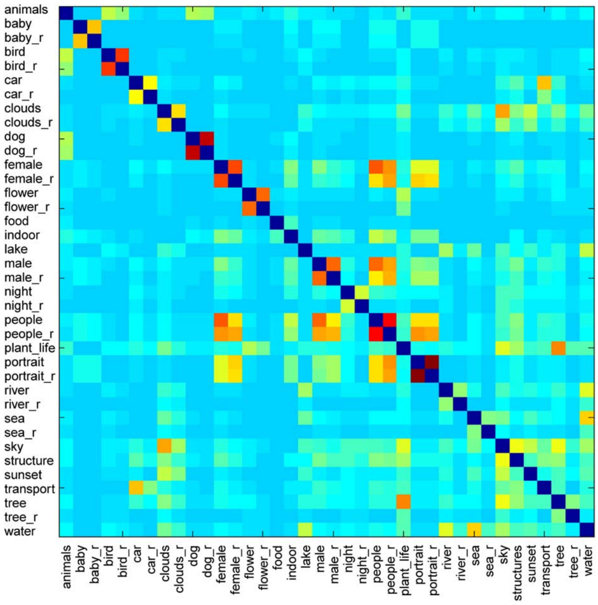


Fig. 5. Pair-wise label correlations for MIR(38) dataset. Labels appear correlatively with each other rather than exist in isolation. Such information is adapted into our graph-based learning framework to improve classification performance.

B. Classification With FMBG

Our first set of experiment results, presented in Fig. 6, compares the performances of multi-label image classification with different graph learning methods using Precision-Recall curves. This includes single modality data graph (e.g. SG_dsift, SG_gist, SG_tags), fused datagraph with all three modalities (FMG), and finally Fused Multimodal Bi-relational Graph (FMBG). In the experiment, each node in the data graph is connected to its top 50 neighbors, and each node in the label graph is connected to its top 5 neighbors. The value of μ in (4) is fixed at 0.1, following the work of [8], [49] for best performance. The value of η in equation (4), which controls the trade-off between data graph and label graph is determined empirically and fixed to 0.11. In terms of initial labels, we randomly selected and labeled 100 images (50 positive, 50 negative) for each class. These parameters are used as default values in our experiment unless otherwise noted. Following typical rank-based classification, images with higher rank are classified as the corresponding class with higher confidence. By varying the confidence threshold, precision values are obtained at different recall levels for each class. The final Precision-Recall curves are computed by taking the average over all classes to show the overall system performance. Notice that the curves are interpolated using standard TREC interpolation rule: for each recall value i from 0.0 to 1.0 with 0.1 increments, we take the maximum precision obtained at any actual recall value greater or equal to i.

From this experiment, we observe that for both datasets, combining graphs from different modalities effectively boosts classification performance. On both MIR(24) and MIR(38) dataset, FMBG improves the performance further by taking into account the label correlation. However, FMBG performs almost the same as FMG (fusion of data graphs) in VOC'07 dataset. This is in fact reasonable, because on average there are only 1.47 labels for each image in the VOC'07 dataset. On the other

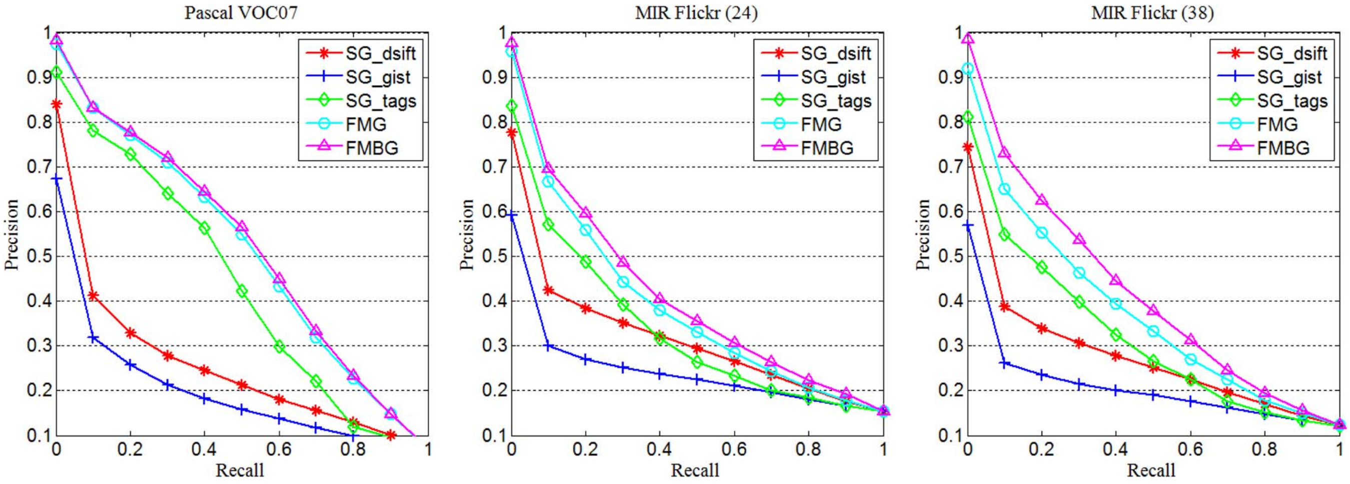


Fig. 6. Precision recall curves for VOC'07, MIR(24) and MIR(38) with different classification methods.

hand, there are 3.72 and 4.72 labels associated with each image in the MIR(24) and MIR(38) dataset respectively, which provide much stronger correlation information among concepts. Hence, the more labels associated the image database, the better FMBG performs due to richer structure of the label graph.

To evaluate the performance of different graph learning methods, we consider their classification on each of the individual classes. In order to reduce the computational load of the experiment, we will only show the results for MIR(38) in subsequent experiments. A common metric to characterize the performance for each class is Average Precision (AP). Basically it computes the area under a non-interpolated Precision-Recall curve. NIST [50] has defined AP as $\sum_{i=1}^{N} p(i)\Delta r(i)$, where N is the total number of images in the collection, $p(i)$ is the precision at cutoff of i images, and $\Delta r(i)$ is the change in recall that happened between cutoff $i-1$ and cutoff i. Mean Average Precision (MAP) is the mean of Average Precision over all classes. Generally speaking, when there are much more negative examples than positive examples, AP serves as a more informative performance measure by revealing a larger gap between classification methods.

Table I shows the AP scores of multi-label image classification for each class. From the results, it is clear that integrating multimodal features can boost classification performance. Overall FMG outperforms the best single modality data graph (Tags) by 6% for multi-label image classification. The improvement on each class is fairly large in magnitude, and they are quite consistent in sign. This also indicates that feature complementariness exists among individual graphs and it can boost classification performance effectively. Further more, by integrating label correlation constraints in the regularization framework, FMBG adds another 5% improvement on top of FMG overall. Out of 38 classes, FMBG outperforms the other methods in 30 classes, by a fairly significant margin. This demonstrate that simultaneously exploiting the consistency in both data graph and label graph is crucial for image classification. To sum up, FMBG gives more than 10% improvement overall compared to the best performing single modality graph in MIR(38).

TABLE I
AP SCORES FOR 38 CLASSES WITH DIFFERENT GRAPH-BASED CLASSIFICATION METHODS. THE BEST RESULT FOR EACH CLASS IS SHOWN IN BOLDFACE

MIR(38)	SG_dsift	SG_gist	SG_tags	FMG	FMBG
animals	0.1996	0.1488	0.3915	0.3383	**0.4322**
baby	0.0221	0.0152	0.2301	0.2724	**0.3614**
baby_r	0.0188	0.0111	0.2339	0.265	**0.3479**
bird	0.0437	0.0398	0.3444	0.2657	**0.3852**
bird_r	0.0319	0.0276	0.3466	0.3621	**0.3949**
car	0.1531	0.0844	0.1683	0.2228	**0.3123**
car_r	0.1589	0.0519	0.3152	0.4285	**0.442**
clouds	0.5559	0.4049	0.3317	0.5597	**0.5825**
clouds_r	0.3772	0.2488	0.1993	**0.444**	0.4313
dog	0.0831	0.0403	0.5736	0.5411	**0.6139**
dog_r	0.0944	0.0466	**0.6416**	0.5607	0.62
female	0.3005	0.2945	0.308	0.3296	**0.3974**
female_r	0.2819	0.2265	0.2086	0.2675	**0.3217**
flower	0.1522	0.0941	**0.3989**	0.3499	0.3974
flower_r	0.1885	0.1334	**0.5238**	0.5043	0.5203
food	0.095	0.0578	0.3388	**0.351**	0.3144
indoor	0.4406	0.3843	0.462	0.4641	**0.4709**
lake	0.1642	0.1064	0.0935	0.1876	**0.2071**
male	0.342	0.3131	0.3105	0.3445	**0.3834**
male_r	0.2373	0.2269	0.2223	0.2669	**0.2998**
night	0.3366	0.1276	0.2644	0.3204	**0.3528**
night_r	0.1679	0.0418	0.1724	**0.2605**	0.2311
people	0.5006	0.5129	0.4868	0.5194	**0.6003**
people_r	0.4616	0.4411	0.4683	0.5133	**0.5808**
plant_life	0.4426	0.3923	0.3995	0.4488	**0.488**
portrait	0.3434	0.2222	0.3252	0.3731	**0.4332**
portrait_r	0.3815	0.286	0.3249	d 0.4251	**0.4742**
river	0.1369	0.0886	0.0811	0.1626	**0.2752**
river_r	0.0285	0.0099	0.014	0.0264	**0.0604**
sea	0.3498	0.2087	0.2347	0.4145	**0.4691**
sea_r	0.1252	0.0469	0.1154	0.2441	**0.2686**
sky	0.5976	0.5217	0.481	0.6321	**0.6757**
structures	0.5662	0.4924	0.4994	0.5698	**0.5743**
sunset	0.3109	0.2352	0.2975	**0.3851**	0.373
transport	0.2501	0.1673	0.1896	0.2571	**0.3409**
tree	0.3706	0.3022	0.2302	0.3446	**0.384**
tree_r	0.1524	0.0768	0.1254	**0.2397**	0.2385
water	0.3886	0.2588	0.2715	0.3848	**0.5274**
Mean	0.2593	0.1944	0.3059	0.3644	**0.4101**

C. Effect of Vertex Degrees in FMBG

An experiment is conducted to identify the effect of vertex degree (graph size) in both data graph and label graph of FMBG. Vertex degree refers to the number of nearest neighbors which connect to a vertex. Intuitively, nodes with very small vertex degree only influence their "local" neighborhoods in the graph diffusion process. This will not help spreading labels, and

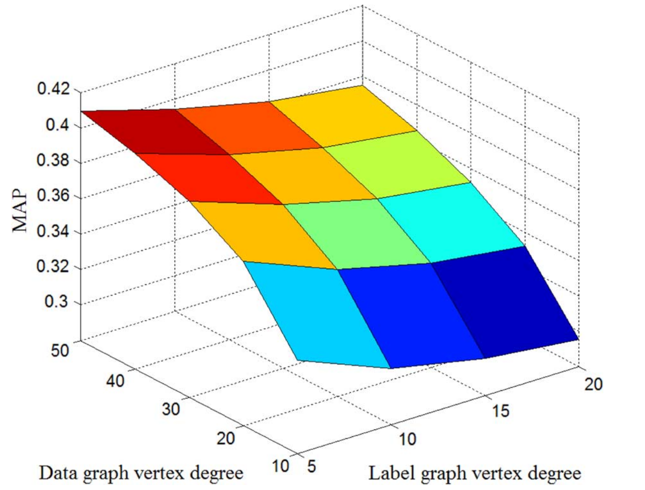


Fig. 7. Performance comparison with different selections of vertex degree in both data graph and label graph.

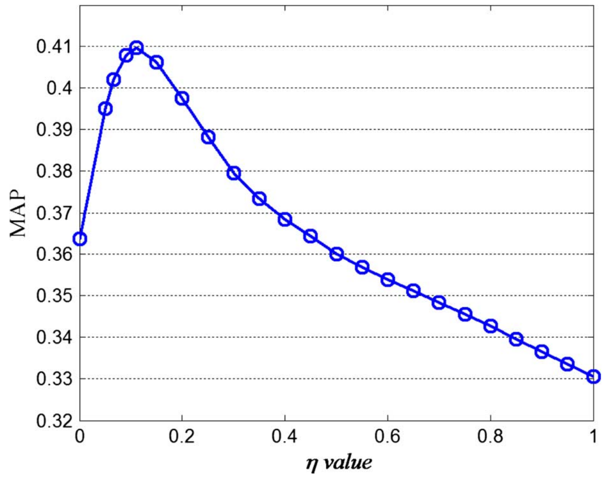


Fig. 8. MAP scores of FBMG with respect to different η values on MIR(38).

subsequently preventing the graph from reaching a "global" stable state. On the other hand, a very large vertex degree may force nodes from different classes to connect together and subsequently suppress diversity. Thus a good leverage on the vertex degree is important for ensuring the performance of graph-based learning.

For this experiment, we keep the same parameters as the previous experiment, with 100 (50 positive, 50 negative) initial labels for each class. We vary the vertex degree for label graph among $\{5, 10, 15, 20\}$ and vary the vertex degree for data graph among $\{10, 20, 30, 40, 50\}$. MAP score of FMBG classification with each pair of combination is shown in Fig. 7. As can be seen, the best performance is obtained when the vertex degree is 5 for the label graph, and 50 for the data graph. The performance decreases as we increase the vertex degree for label graph. There is about 3% drop in terms of MAP score when changing the vertex degree from 5 to 10 alone. This is probably because there is only a small number of classes in the dataset, larger vertex degree introduces unreliable links regarding label correlation. On the other hand, the performance improves as we gradually increase the vertex degree in data graph from 10 to 50. The improvement on MAP scores is about 7%. This is a good example of a overly small graph size prevent information propagating from one node to another. In some extreme cases, it may even disconnect part of the graph. A decent graph size seems to be 50 in this case. Although not shown in Fig. 7, the performance is almost saturated around that point. The improvement in terms of MAP score is not significant beyond that. In all subsequent experiments, we will fix the vertex degree for data graph at 50, and vertex degree for label graph at 5.

D. Trade-Off Between Data Graph and Label Graph

In the third experiment, we investigate the trade-off between the two sub-graphs in FMBG, namely fused data graph and the label graph. These two sub-graphs are reflected by the two regularizers in (4). The trade-off between the two is controlled by the parameters η. When η is set to zero, the regularizer for label correlation is completely ignored, and our framework is reduced to multimodal graph learning. As we increase η value, the emphasis on label correlation also increases. Intuitively higher weight should be assigned to the fused data graph since it is significantly larger and filled with richer structure information. In addition, multiple evidences are aggregated together to estimate the affinities among vertices in the fused data graph.

The same classification task is carried out with FMBG using different η values. Experimental results in Fig. 8 shows the MAP scores with respect to η. As we can see, the result actually matches with our initial guess. The best MAP score is obtained when $\eta = 0.11$, which means the regularizer for label graph is weighed about 1/10 of the weight assigned to the fused data graph. Performance decrease as we increase η beyond that point. This is because label correlation is overly emphasized while it can not provide any more discriminant power; on the other hand the richer information from fused data graph is neglected. By choosing the appropriate η value, our method effectively leverages on the two sub-graphs and improves the performance compared with methods which only operate within one sub-graph (when $\eta = 0$).

E. Effect of Different Size of Labeled Data

For this experiment, we investigate the multi-label classification performance with respect to different sizes of initial label sets. We perform the same classification task with different graph-based learning methods, and gradually increase the initial labeled set size from 10 to 100, For each size we randomly selected a set of images for initial labeling. Intuitively, a better performance is expected with a larger initial set of labels.

The results are shown on Fig. 9. The MAP scores of all methods increase as the number of initial labels increases. However, the gain on MAP scores varies, with FMBG being the most significant (16%), followed by FMG (12%). The gain for single modality graphs are much less (varies from 3% to 7%), especially with SG_gist as its gain on MAP score saturates toward the end. This again indicates the importance of data fusion since it is very hard to achieve satisfactory results with any single feature alone. Overall, our proposed method consistently outperforms all other methods throughout the experiment.

F. Evaluation of Feature Combination

In the final experiment, we investigate the image classification performance with different feature fusion strategies. We

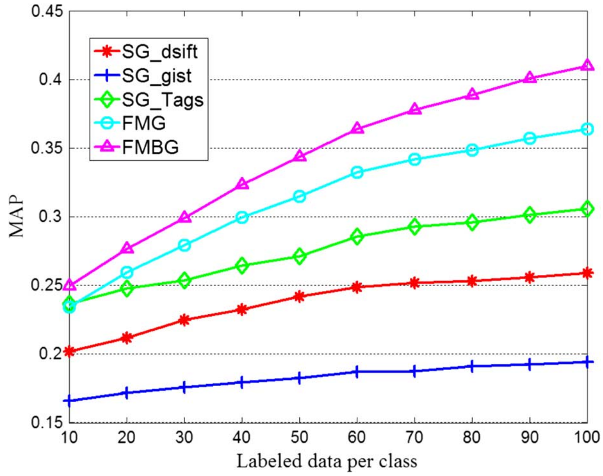


Fig. 9. Performance variation with respect to different number of initial labels on MIR(38).

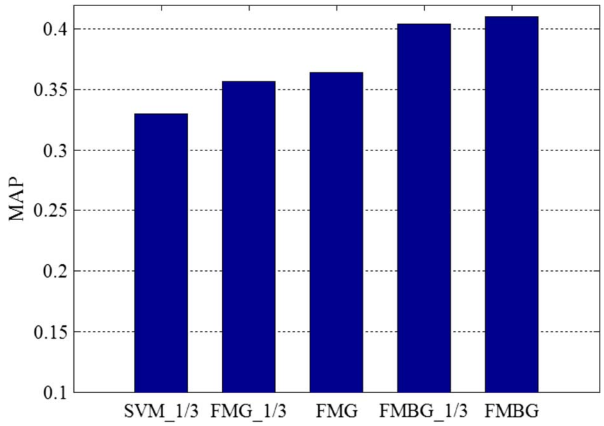


Fig. 10. MAP scores with respect to different feature combination mechanisms.

start with training a baseline SVM classifier using multiple features. To do that, all feature vectors are first normalized to have L2-norm of 1, then features of different types are concatenated together to form a long vector. A linear SVM is trained for each class based on the concatenated feature vector. The training samples are the same set of images used for initial labels in the graph-based classification approach. In terms of graph-based fusion, the naive approach is putting equal weights on individual graph and then doing a "union" sum to combine them. A more sophisticated way is to learn the weight α_g for each graph iteratively before combining them, as we do in our framework for FMG and FMBG. Thus we also evaluate the performance improvement of learning the weight parameters α_g for feature fusion.

Results are shown in Fig. 10. SVM_1/3 stands for the standard feature concatenation using SVM; FMG_1/3 and FMBG_1/3 stand for the graph-based fusion with equal weights; and FMG and FMBG are the graph-based fusion with learnt weights. Overall, graph-based methods outperform SVM. This is because all the graph-based methods exploit the structure of unlabeled data, which is beneficial when initial training data is not large. In addition, incorporating graphs with varying weights does gives better results than using all graph with equal weights. However, the improvement does not seem very significant, and not as promising as shown in [43]. Recall that the weight α_g is affected by r as shown in (7). The values of r is tuned per class in the work of [43] for binary (single label) classification. However, this is not possible in our case when doing multi-label classification. We fix $r = 1.1$ in our experiment for best performance. The bottom line is feature combination is important for classification, and using graph smoothness to derive weights for each feature is a valid choice to deliver better performance.

V. Conclusions

This work addresses the multi-label image classification problem by jointly considering the complementary nature of multimodal features and correlated nature of labels. We formulate the problem as a semi-supervised label diffusion process on a unified bi-relational graph. Such a representation enables effective fusion of multiple features to propagate a sparse set of initial labels over the entire image dataset. Extensive experimental results demonstrate the superiority of the proposed method over methods that treat label correlations of feature fusion in isolation.

Acknowledgment

The authors would like to thank Dr. Ziyu Guan for his helpful discussions.

References

[1] M. R. Boutell, J. Luo, X. Shen, and C. M. Brown, "Learning multi-label scene classification," *Pattern Recognit.*, vol. 37, no. 9, 2004.

[2] S. Zhu, X. Ji, W. Xu, and Y. Gong, "Multi-labelled classification using maximum entropy method," in *Proc. ACM SIGIR*, 2005, pp. 274–281.

[3] Y. Yang, F. Nie, D. Xu, J. Luo, Y. Zhuang, and Y. Pan, "A multimedia retrieval framework based on semi-supervised ranking and relevance feedback," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 4, pp. 723–742, Apr. 2012.

[4] , O. Chapelle, B. Schölkopf, and A. Zien, Eds.*, Semi-Supervised Learning*. Cambridge, MA, USA: MIT Press, 2006.

[5] V. Vapnik*, Statistical Learning Theory*. New York, NY, USA: Wiley, 1998.

[6] K. Nigam, A. K. McCallum, S. Thrun, and T. Mitchell, "Text classification from labeled and unlabeled documents using em," *Mach. Learn.*, vol. 39, pp. 103–134, 2000.

[7] X. Zhu, Z. Ghahramani, and J. D. Lafferty, "Semi-supervised learning using Gaussian fields and harmonic functions," in *Proc. ICML*, 2003.

[8] D. Zhou, O. Bousquet, T. N. Lal, J. Weston, and B. Schölkopf, "Learning with local and global consistency," in *Proc. NIPS*, Cambridge, MA, USA: MIT Press, 2003.

[9] J. H. He, M. Li, H. Zhang, H. Tong, and C. Zhang, "Manifold-ranking based image retrieval," in *Proc. ACM Multimedia*, 2004, pp. 9–16.

[10] C. Wang, F. Jing, L. Zhang, and H. Zhang, "Image annotation refinement using random walk with restarts," in *Proc. ACM Multimedia*, 2006.

[11] J. Tang, S. Yan, R. Hong, G.-J. Qi, and T.-S. Chua, "Inferring semantic concepts from community-contributed images and noisy tags," in *Proc. ACM Multimedia*, 2009, pp. 223–232.

[12] M. Guillaumin, T. Mensink, J. J. Verbeek, and C. Schmid, "Tagprop: Discriminative metric learning in nearest neighbor models for image auto-annotation," in *Proc. ICCV*, 2009, pp. 309–316.

[13] H.-K. Tan and C.-W. Ngo, "Fusing heterogeneous modalities for video and image re-ranking," in *Proc. ICMR*, 2011, p. 15.

[14] Z. Fu, H. H. Ip, H. Lu, and Z. Lu, "Multi-modal constraint propagation for heterogeneous image clustering," in *Proc. ACM Multimedia*, 2011.

[15] C. Ji, X. Zhou, L. Lin, and W. Yang, "Labeling images by integrating sparse multiple distance learning and semantic context modeling," in *Proc. ECCV*, 2012, pp. 688–701.

[16] S. Clinchant, J. Ah-Pine, and G. Csurka, "Semantic combination of textual and visual information in multimedia retrieval," in *Proc. ICMR*, 2011.

[17] M. Wang, X.-S. Hua, X. Yuan, Y. Song, and L.-R. Dai, "Optimizing multi-graph learning: Towards a unified video annotation scheme," in *Proc. ACM Multimedia*, 2007, pp. 862–871.

[18] J. Tang, H. Li, G.-J. Qi, and T.-S. Chua, "Image annotation by graph-based inference with integrated multiple/single instance representations," *IEEE Trans. Multimedia*, vol. 12, no. 2, pp. 131–141, Feb. 2010.

[19] E. Moxley, T. Mei, and B. S. Manjunath, "Video annotation through search and graph reinforcement mining," *IEEE Trans. Multimedia*, vol. 12, no. 3, pp. 184–193, Apr. 2010.

[20] A. Elisseeff and J. Weston, "A kernel method for multi-labelled classification," in *Proc. NIPS*, Cambridge, MA, USA: MIT Press, 2001, pp. 681–687.

[21] S. Godbole and S. Sarawagi, "Discriminative methods for multi-labeled classification," in *Proc. 8th Pacific-Asia Conf. Knowledge Discovery and Data Mining*, New York, NY, USA: Springer, 2004, pp. 22–30.

[22] S. Kumar and M. Hebert, "Discriminative fields for modeling spatial dependencies in natural images," in *Proc. NIPS*, Cambridge, MA, USA: MIT Press, 2003.

[23] Y. Liu, R. Jin, and L. Yang, "Semi-supervised multi-label learning by constrained non-negative matrix factorization," in *Proc. AAAI*, 2006.

[24] F. Kang, R. Jin, and R. Sukthankar, "Correlated label propagation with application to multi-label learning," in *Proc. CVPR*, 2006, pp. 1719–1726.

[25] G.-J. Qi, X.-S. Hua, Y. Rui, J. Tang, T. Mei, and H.-J. Zhang, "Correlative multi-label video annotation," in *Proc. ACM Multimedia*, 2007.

[26] J. Wang, Y. Zhao, X. Wu, and X.-S. Hua, "Transductive multi-label learning for video concept detection," in *Proc. Multimedia Information Retrieval*, 2008, pp. 298–304.

[27] Y. Xiang, X. Zhou, T.-S. Chua, and C.-W. Ngo, "A revisit of generative model for automatic image annotation using Markov random fields," in *Proc. CVPR*, 2009, pp. 1153–1160.

[28] J. Tang, X.-S. Hua, M. Wang, Z. Gu, G.-J. Qi, and X. Wu, "Correlative linear neighborhood propagation for video annotation," *IEEE Trans. Syst., Man, Cybern. B*, vol. 39, no. 2, pp. 409–416, 2009.

[29] S. Ji, L. Tang, S. Yu, and J. Ye, "A shared-subspace learning framework for multi-label classification," *ACM Trans. Knowl. Discov. Data*, vol. 4, no. 2, May 2010, article no. 8.

[30] K. Yu, S. Yu, and V. Tresp, "Multi-label informed latent semantic indexing," in *Proc. ACM SIGIR Research and Development in Information Retrieval*, 2005, pp. 258–265.

[31] H. Wang, H. Huang, and C. Ding, "Image annotation using bi-relational graph of images and semantic labels," in *Proc. CVPR*, 2011, pp. 793–800.

[32] G.-X. Yu, C. Domeniconi, H. Rangwala, G. Zhang, and Z. Yu, "Transductive multi-label ensemble classification for protein function prediction," in *Proc. KDD*, 2012, pp. 1077–1085.

[33] H. Wang, H. Huang, and C. H. Q. Ding, "Image annotation using multi-label correlated green's function," in *Proc. IEEE ICCV*, 2009, pp. 2029–2034.

[34] G. Chen, Y. Song, F. Wang, and C. Zhang, "Semi-supervised multi-label learning by solving a Sylvester equation," in *Proc. SDM*, 2008, pp. 410–419.

[35] Z.-J. Zha, T. Mei, J. Wang, Z. Wang, and X.-S. Hua, "Graph-based semi-supervised learning with multiple labels," *J. Vis. Commun. Image Represent.*, vol. 20, no. 2, pp. 97–103, 2009.

[36] H. Wang, H. Huang, and C. H. Q. Ding, "Discriminant Laplacian embedding," in *Proc. AAAI*, 2010.

[37] M. Varma and D. Ray, "Learning the discriminative power-invariance trade-off," in *Proc. ICCV*, 2007, pp. 1–8.

[38] A. Vedaldi, V. Gulshan, M. Varma, and A. Zisserman, "Multiple kernels for object detection," in *Proc. ICCV*, 2009, pp. 606–613.

[39] M. Guillaumin, J. Verbeek, and C. Schmid, "Multimodal semi-supervised learning for image classification," in *Proc. CVPR*, 2010, pp. 902–909.

[40] L. Bo, K. Lai, X. Ren, and D. Fox, "Object recognition with hierarchical kernel descriptors," in *Proc. CVPR*, 2011, pp. 1729–1736.

[41] L. Bo, X. Ren, and D. Fox, "Depth kernel descriptors for object recognition," in *Proc. IROS*, 2011, pp. 821–826.

[42] M. Belkin and P. Niyogi, "Semi-supervised learning on Riemannian manifolds," *Mach. Learn.*, vol. 56, pp. 209–239, 2004.

[43] M. Wang, X.-S. Hua, R. Hong, J. Tang, G.-J. Qi, and Y. Song, "Unified video annotation via multigraph learning," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 19, no. 5, pp. 733–746, 2009.

[44] E. Anderson, Z. Bai, C. Bischof, S. Blackford, J. Demmel, J. Dongarra, J. Du Croz, A. Greenbaum, S. Hammarling, A. McKenney, and D. Sorensen*, LAPACK Users' Guide*, 3rd ed. Philadelphia, PA, USA: SIAM, 1999.

[45] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, The PASCAL Visual Object Classes Challenge 2007 (VOC2007).

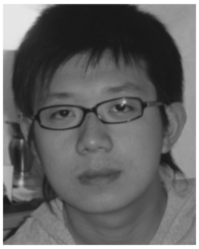
[46] M. J. Huiskes and M. S. Lew, "The mir flickr retrieval evaluation," in *Proc. 2008 ACM Int. Conf. Multimedia Information Retrieval (MIR'08)*, New York, NY, USA: ACM, 2008.

[47] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

[48] B. Manjunath and W. Ma, Texture Features for Browsing and Retrieval of Image Data (Derivation of the Gabor Filter Dictionary Parameters), Tech. Rep. 8, Aug. 1996.

[49] D. Zhou, J. H. Huang, and B. Scholkopf, "Learning from labeled and unlabeled data on a directed graph," in *Proc. ICML*, 2005, pp. 1036–1043.

[50] N. I. of Standards and Technology [Online]. Available: http://www.nist.gov/index.html



Jiejun Xu received his M.S. and Ph.D. degree in computer science from University of California Santa Barbara in 2007 and 2013, respectively. His research interests are computer vision and multimedia retrieval. Prior to joining the Ph.D. program, he was with the Center for Bio-Image Informatics, and worked on developing the Bio-Image Semantic Query User Environment (Bisque). He has interned with Telefonia Research in Barcelona and HRL Laboratories in Malibu. He was also a recipient of the NSF IGERT fellowship in 2007.

Vignesh Jagadeesh received his Ph.D. in Electrical Engineering from UCSB in 2013. He holds an M.S. in ECE from UCSB, and a B.S. from Anna University (India). His research interests span computer vision, image analysis and processing. His graduate research work is on interactive and scalable image analysis to handle large scale image mosaics generated by high throughput electron microscopy. He has interned for two summers with True Vision 3D Surgical, one summer each with Mayachitra Inc. and eBay Research.

B. S. Manjunath (F'05) received the B.E. degree (with distinction) in electronics from Bangalore University, Bangalore, India, in 1985, the M.E. degree (with distinction) in systems science and automation from the Indian Institute of Science, Bangalore, in 1987, and the Ph.D. degree in electrical engineering from University of Southern California, Los Angeles, in 1991. He is now a Professor of electrical and computer engineering and Director of the Center for Bio-Image Informatics at the University of California, Santa Barbara. His current research interests include image processing, data hiding, multimedia databases, and bio-image informatics. He has published over 250 peer-reviewed articles on these topics and is a co-editor of the book Introduction to MPEG-7 (Wiley, 2002). He was an associate editor of the IEEE TRANSACTIONS ON IMAGE PROCESSING, PATTERN ANALYSIS AND MACHINE INTELLIGENCE, MULTIMEDIA, INFORMATION FORENSICS, the IEEE SIGNAL PROCESSING LETTERS and is currently an AE for the BMC Bioinformatics Journal. He is a co-author of the paper that won the 2013 Transactions on Multimedia best paper award and is a fellow of the IEEE.