Graph-based Semi-Supervised Learning with Multiple Labels

Zheng-Jun Zha, Tao Mei, Jingdong Wang, Zengfu Wang, Xian-Sheng Hua

PII: S1047-3203(08)00114-4
DOI: 10.1016/j.jvcir.2008.11.009
Reference: YJVCI 790

To appear in: *J. Vis. Commun. Image R.*

Received Date: 2 June 2008
Revised Date: 12 October 2008
Accepted Date: 12 November 2008

Please cite this article as: Z-J. Zha, T. Mei, J. Wang, Z. Wang, X-S. Hua, Graph-based Semi-Supervised Learning with Multiple Labels, *J. Vis. Commun. Image R.* (2008), doi: 10.1016/j.jvcir.2008.11.009

Graph-based Semi-Supervised Learning with Multiple Labels

Zheng-Jun Zha[†] [‡] Tao Mei[*] Jingdong Wang[*] Zengfu Wang[‡] Xian-Sheng Hua[*]

[†] *MOE-Microsoft Key Laboratory of Multimedia Computing and Communication, University of Science and Technology of China, Hefei, 230027, P. R. China*

[‡] *Department of Automation, University of Science and Technology of China, Hefei, 230027, P. R. China*

[*] *Internet Media Group, Microsoft Research Asia, Beijing, 100190, P. R. China*

Abstract

Conventional graph-based semi-supervised learning methods predominantly focus on single label problem. However, it is more popular in real-world applications that an example is associated with multiple labels simultaneously. In this paper, we propose a novel graph-based learning framework in the setting of semi-supervised learning with multiple labels. This framework is characterized by simultaneously exploiting the inherent correlations among multiple labels and the label consistency over the graph. Based on the proposed framework, we further develop two novel graph-based algorithms. We apply the proposed methods to video concept detection over TRECVID 2006 corpus and report superior performance compared to the state-of-the-art graph-based approaches and the representative semi-supervised multi-label learning methods.

Key words: graph-based learning, multiple labels, semi-supervised learning, video concept detection

1. Introduction

In real-world applications of machine learning, a large amount of unlabeled data is available, while it is costly to obtain labeled data since human labeling is a labor-intensive and time-consuming process. For example, in video concept detection, one may have an easy access to a large database of videos, but a small part of them is manually annotated.

Consequently, semi-supervised learning, which attempts to leverage the unlabeled data in addition to labeled data, has attracted much attention. Many different semi-supervised learning techniques have been proposed. These methods include, among others, EM with generative mixture models [22], co-training [19], self training [21], and transductive support vector machines [20]. Extensive review can be found in [2].

In recent years, the most active area of research in semi-supervised learning has been in graph-based semi-supervised learning methods [3] [4] [5] [10] [11] [12], which model the whole data set as a graph such that the nodes correspond to labeled and unlabeled data points, and the edges reflect the similarities between data points. Almost all the graph-based methods essentially estimate a function on the graph such that it has two properties: 1) it should be close to the given labels on the labeled examples, and 2) it should be smooth on the whole graph. This can be expressed in a regularization framework where the first term is a loss function to penalize the deviation from the given labels, and the second term is a regularizer to prefer the label smoothness. The typical graph-based methods are similar to each other, and differ slightly in the loss function and the regularizer.

Although the graph-based approaches have been proved effective, they mainly limit in dealing with single label problems. However, many real-world tasks are naturally posed as multi-label problems, where an example can be associated with multiple labels simultaneously. For example, in video concept detection, a video clip can be annotated with multiple labels at the same time, such as "sky," "mountain," and "water" (see Fig. 1). Similarly, in text categorization, a document can be classified into multiple categories, e.g., a document can belong to "novel," "Jules Verne's writing," and "books on travelling." A direct way to tackle the typical graph-based learning under multi-label setting is to translate it into a set of independent single label problems [6]. That is to say, the multi-label problems need to be implemented label-by-label independently. The drawback

Email address: zzjun@mail.ustc.edu.cn, tmei@microsoft.com, i-jingdw@microsoft.com, zfwang@ustc.edu.cn and xshua@microsoft.com (Zheng-Jun Zha[†] [‡] Tao Mei[*] Jingdong Wang[*] Zengfu Wang[‡] Xian-Sheng Hua[*]).

Fig. 1. Sample video clips associated with multiple labels.

with this strategy is that it does not take into account the inherent correlations among multiple labels. However, the labels are usually interacting with each other naturally. For example, "mountain" and "sky" tend to appear simultaneously, while "sky" typically does not appear with "indoor." The value of label correlations has been proven in various application fields [7] [8] [9] [31] [14] [26].

To address the above issue, in this paper, we propose a novel graph-based semi-supervised learning framework, which can simultaneously explore the correlations among multiple labels and the label consistency over the graph. The vector-valued function estimated on the graph has three properties: 1) it should be close to the given labels, 2) it should be smooth on the whole graph, and 3) it should be consistent with the label correlations. Specifically, the framework employs two types of regularizers. One is used to prefer the label smoothness on the graph, the other is adopted to address that the multi-label assignments for each example should be consistent with the inherent label correlations.

Based on this framework, we develop two novel graph-based learning algorithms, Multi-label gaussian random field (ML-GRF) and Multi-label local and global consistency (ML-LGC) algorithm, which are the extensions of two existing graph-based methods: the Gaussian random field and Harmonic function (GRF) [4] method and the Local and global consistency (LGC) method [3], respectively. The proposed algorithms are shown to be able to make use of both labeled and unlabeled data, and the correlations among labels. We apply these two algorithms to video concept detection and conduct experiments over TRECVID 2006 development set [15]. We report the superior performance compared to key existing graph-based learning approaches and semi-supervised multi-label learning methods.

To summarize, the main contributions of this work include:

– It provides a novel graph-based learning framework to address the multi-label problems, which is an extension of the existing graph-based framework.
– It develops two graph-based algorithms (i.e., ML-GRF and ML-LGC) under the proposed framework. Compared to the typical graph-based methods, ML-GRF and ML-LGC can simultaneously exploit both the labeled and unlabeled data, and the correlations among multiple labels .
– We have performed extensive experimental comparisons of the proposed algorithms, the representative graph-based approaches, and existing semi-supervised multi-label algorithms.

Compared to our preliminary work [13], in this paper, we comprehensively analyze the multiple label phenomena and provide deeper evaluation. The rest of this paper is organized as follows. We firstly give a brief summary of related work on semi-supervised learning and multi-label learning in Section 2. Section 3 mathematically introduces the existing graph-based learning framework. Section 4 provides the detailed description of the proposed framework and algorithms. Experimental results on TRECVID 2006 data set are reported in Section 5 followed by concluding remarks in Section 6.

2. Related Work

2.1. *Semi-Supervised Learning*

In this subsection, we only review the intimately related work, i.e., graph-based semi-supervised learning methods. More discussion of semi-supervised learning can be found in [2].

The graph-based learning strategy models the whole data set as a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. $\mathcal{V}$ is the vertex set, which is equivalent to the data set, and $\mathcal{E}$ is the set of edges. A nonnegative weight w_{ij} is associated with each edge e_{ij} and w_{ij} reflects how similar datum i is to datum j. Many graph-based methods are based on the common assumption that the labels are smooth on the graph. They aim at estimating a function f over the graph such that it simultaneously satisfies two properties: 1) it should be close to the given labels on the labeled nodes, and 2) it should be smooth on the whole graph. This can be formulated in a regularization framework where the first term is a loss function, and the second term is a regularizer. The graph-based methods are similar to each other, and differ slightly in the definition of the loss function and the regularizer. For example, Zhu *et al.* [4] proposed a Gaussian random field and Harmonic function (GRF) method. They defined a quadratic loss function with infinity weight to clamp the labeled examples, and formulated the regularizer based on

the graph combinatorial Laplacian. Belkin *et al* [5] mentioned that p-Laplacian can be used as a regularizer. In [3], Zhou *et al.* presented the Local and global consistency (LGC) method. They also defined a quadratic loss function and used the normalized combinatorial Laplacian in the regularizer. Since the graph is at the heart of graph-based methods, there also exist some works related to the issue of graph construction. For example, the widely-used graph is K-nearest-neighbor (K-NN) graph, where each node is connected to its K nearest neighbors under some distance measure and the edges can be weighted by a Gaussian function $w_{ij} = exp(-||x_i - x_j||^2/\sigma^2)$, or unweighted ($w_{ij} = 1$). Recently, Tang *et al.* [10] [11] and Wang *et al.* [12] proposed structure-sensitive graph-based methods. They addressed that the similarities among the samples are not merely related to distances but also related to the structures around the samples.

2.2. *Multi-Label Learning*

Multi-label learning problems widely exist in real-world applications, such as text classification [7] [14] [24] [27] [29], functional genomic [25], image classification [28] [26], and video concept detection [9] [31].

The typical solution of multi-label learning is to translate the multi-label learning task into a set of single-label problems. For example, Boutell *et al.* [28] solved the multi-label scene classification problem by building individual classifier for each label. The labels of a new sample are determined by the outputs of these individual classifiers. Such solution treats the labels in isolation and ignores the correlations among the labels. To exploit label correlations, some researchers have proposed fusion-based methods [27]. Godbole *et al.* [27] proposed to leverage the correlations by adding a contextual fusion step based on the outputs of the individual classifiers. More sophisticated multi-label learning approaches model labels and correlations between labels simultaneously [14] [26] [9]. Liu *et al.* [31] [14] proposed a multi-label learning method based on constrained nonnegative matrix factorization. Kang *et al.* [26] developed a *Correlation Label Propagation* (CLP) approach to explicitly capture the interactions between labels. Rather than treating labels independently, CLP simultaneously co-propagates multiple labels from training examples to testing examples. More recently, Qi *et al.* [9] proposed a unified *Correlative Multi-Label* (CML) *Support Vector Machine* (SVM) to simultaneously classify labels and model their correlations in a new feature space which encodes both the label models and their interactions together. Chen and Hauptmann [31] proposed multi-concept discriminative random field (MDRF) to automatically identify concept correlations and learn concept classifiers simultaneously. These two methods are similar in spirit as they modify the regularization term of support vector machines (in [9]) or logistic regression (in [31]) to accommodate the correlations between different concepts.

3. Graph-based Semi-supervised Learning with Single Label

As aforementioned, most graph-based methods are based on the common assumption that the labels are smooth on the graph. Then, they essentially estimate a function over the graph such that it satisfies two conditions: 1) it should be close to the given labels, and 2) it should be smooth on the whole graph. Generally, these two conditions are presented in a regularization form.

Let $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N\}$ be a set of N data points in $\mathbb{R}^d$. The first L points are labeled as $\mathbf{y}_l = [y_1, y_2, \cdots, y_L]^T$ with $y_i \in \{0, 1\}$ $(1 \leqslant i \leqslant L)$, and the task is to label the remaining points $\{\mathbf{x}_{L+1}, \cdots, \mathbf{x}_N\}$. Denote the graph by $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where the node set $\mathcal{V} = \mathcal{L} \cup \mathcal{U}$ with $\mathcal{L}$ corresponding to $\{\mathbf{x}_1, \cdots, \mathbf{x}_L\}$ and $\mathcal{U}$ corresponding to $\{\mathbf{x}_{L+1}, \cdots, \mathbf{x}_N\}$. The edges $\mathcal{E}$ are weighted by the $N \times N$ affinity matrix $\mathbf{W}$ with w_{ij} indicating the similarity measure between $\mathbf{x}_i$ and $\mathbf{x}_j$, and w_{ii} is set to 0. Let $\mathbf{f} = [f_1, f_2, \cdots, f_L, f_{L+1}, \cdots, f_N]^T$ denote the predicted labels of $\mathcal{X}$.

Mathematically, the graph-based methods aim to find an optimal $\mathbf{f}^*$ essentially by minimizing the following objective function

$$E(\mathbf{f}) = E_l(\mathbf{f}) + E_s(\mathbf{f}), \tag{1}$$

where $E_l(\mathbf{f})$ is a loss function to penalize the deviation from the given labels, and $E_s(\mathbf{f})$ is a regularizer to prefer the label smoothness. For example, the Gaussian random field method [4] formulates $E_l(\mathbf{f})$ and $E_s(\mathbf{f})$ as

$$E_l(\mathbf{f}) = \infty \sum\nolimits_{i \in L} (f_i - y_i)^2 = (\mathbf{f} - \mathbf{y})^T \mathbf{\Lambda} (\mathbf{f} - \mathbf{y}),$$
$$E_s(\mathbf{f}) = \frac{1}{2} \sum\nolimits_{i,j \in L \cup U} w_{ij} (f_i - f_j)^2 = \mathbf{f}^T \Delta \mathbf{f},$$

where $\mathbf{\Lambda}$ is a diagonal matrix with $\mathbf{\Lambda}_{ii} = \infty$, $i \leqslant L$ and $\mathbf{\Lambda}_{ii} = 0$, $i > L$. $\Delta = \mathbf{D} - \mathbf{W}$ is the combinatorial graph Laplacian, $\mathbf{D}$ is a diagonal matrix with its (i, i)-element equal to the sum of the ith row of $\mathbf{W}$.

In Local and Global Consistency method [3], a similar graph-based method is developed. Specifically, $E_l(\mathbf{f})$ and $E_s(\mathbf{f})$ are defined as

$$E_l(\mathbf{f}) = \mu \sum\nolimits_{i \in L \cup U} (f_i - y_i)^2 = (\mathbf{f} - \mathbf{y})^T (\mathbf{f} - \mathbf{y}),$$
$$E_s(\mathbf{f}) = \frac{1}{2} \sum\nolimits_{i,j \in L \cup U} w_{ij} (f_i/\sqrt{d_i} - f_j/\sqrt{d_j})^2 = \mathbf{f}^T \mathbf{D^{-1/2}} \Delta \mathbf{D^{-1/2}} \mathbf{f},$$

where $\mathbf{D}^{-1/2} \Delta \mathbf{D}^{-1/2}$ is the normalized combinatorial Laplacian.

The existing graph-based methods mainly address the semi-supervised problem for single label scenario, and they are sub-optimal for multi-label scenario, which is more challenging but much closer to real-world applications.

4. Graph-based Semi-Supervised Learning with Multiple Labels

In this section, we address the semi-supervised K-label problem. Define an $N \times K$ label matrix $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_N]^T$, where y_{ik} is 1 if $\mathbf{x}_i$ is a labeled sample with its label k, and 0 otherwise. Define an $N \times K$ matrix $\mathbf{F} = [\mathbf{f}_1, \mathbf{f}_2, \cdots, \mathbf{f}_N]^T$, where $\mathbf{f}_i = [f_{i1}, f_{i2}, \cdots, f_{iK}]^T$ is a K-dimensional label vector. f_{ik} is the confidence that $\mathbf{x}_i$ is associated with label k.

4.1. *Motivation*

In multi-label scenario, it is observed that the labels assigned to a sample (e.g., a video clip in video concept detection task) are usually consistent with the inherent label correlations. For example, "car," and "road" are usually co-assigned to a certain video clip since they often appear simultaneously, while "explosion fire" and "waterscape" are generally not assigned to a sample at the same time since they usually do not co-occur.

Motivated by this observation, we propose a novel graph-based learning framework, such that the vector-valued function over the graph has three properties: 1) it should be close to the given labels, 2) it should be smooth on the whole graph, and 3) it should be consistent with the label correlations. The former two properties have been addressed in existing graph-based methods. The third one is novel and the key contribution of this paper. Recently, a similar semi-supervised learning framework was developed in a parallel work [1]. In [1], Chen *et al.* proposed to address the multi-label problems by extending the existing graph-based methods. Specifically, a category level graph was incorporated into the existing methods. The category level graph was defined on all the categories, where each node represents each category and the edge weight reflects the similarity between two categories. They defined a regularization term to address the smoothness of the labels of categories. It can be found that Chen's and our methods are similar in spirit.

4.2. *Formulation*

4.2.1. *Framework*

This unified regularization framework consists of three components: a loss function $E_l(\mathbf{F})$, and two types of regularizers $E_s(\mathbf{F})$ and $E_c(\mathbf{F})$. Specifically, $E_l(\mathbf{F})$ corresponds to the first property to penalize the deviation from the given multi-label assignments, $E_s(\mathbf{F})$ is a regularizer to address the label smoothness, and $E_c(\mathbf{F})$ is a regularizer to prefer the third property. Then, the proposed framework can be formulated to minimize

$$E(\mathbf{F}) = E_l(\mathbf{F}) + E_s(\mathbf{F}) + E_c(\mathbf{F}), \tag{2}$$

where $E_l(\mathbf{F})$ and $E_s(\mathbf{F})$ can be specified in a way similar to that adopted in existing graph-based methods.

We make use of the correlations among multiple labels to define $E_c(\mathbf{F})$. To capture the label correlations, we introduce a $K \times K$ symmetric matrix $\mathbf{C}$ with c_{ij} represents the correlation between label i and label j. Then, given a label matrix $\mathbf{F}_c = [\mathbf{f}_1\ \mathbf{f}_2\ \cdots\ \mathbf{f}_n]^T$ on n data points with certain labeling, $\mathbf{C}^{'}$ is calculated as $c^{'}_{ij} = \exp(-||\mathbf{f}^{'}_i - \mathbf{f}^{'}_j||^2/2\sigma_c^2)$, where $\mathbf{f}^{'}_i$ is the ith column of $\mathbf{F}_c$, and $\sigma_c = \mathrm{E}(\parallel \mathbf{f}^{'}_i - \mathbf{f}^{'}_j \parallel)$ is the average distance. Then $\mathbf{C}$ is defined as $\mathbf{C} = \mathbf{C}^{'} - \mathbf{D}_c$, where $\mathbf{D}_c$ is a diagonal matrix with the (i,i)-element equal to the sum of the ith row of $\mathbf{C}^{'}$. Let us define e_i as $\mathbf{f}_i^T\mathbf{C}\mathbf{f}_i$. Intuitively, e_i reflects the coherence between the inherent correlation and the label vector $\mathbf{f}_i$ assigned to $\mathbf{x}_i$. That is to say, the larger e_i is, $\mathbf{f}_i$ is more coherent with the label correlations. Consequently, we specify $E_c(\mathbf{F}) = -\mathrm{tr}(\mathbf{F}\mathbf{C}\mathbf{F}^T)$ to make the predicted multiple labels for each sample satisfy the correlations, where $\mathrm{tr}(\mathbf{M})$ is the trace of matrix $\mathbf{M}$.

4.2.2. *Algorithm instances: ML-GRF and ML-LGC*

The proposed framework provides us a powerful platform to design novel algorithms. Here we describe two algorithm instances developed under this framework.

Firstly, we define $E_l(\mathbf{F})$ and $E_s(\mathbf{F})$ as the same as those in GRF [4], i.e.,

$$E_l(\mathbf{F}) = \mathrm{tr}((\mathbf{F} - \mathbf{Y})^T\mathbf{\Lambda}(\mathbf{F} - \mathbf{Y})),$$
$$E_s(\mathbf{F}) = \mathrm{tr}(\mathbf{F}^T\Delta\mathbf{F}).$$

Eq. (2) is then specifically formulated as the following

$$E(\mathbf{F}) = \mathrm{tr}((\mathbf{F}-\mathbf{Y})^{\mathbf{T}}\Lambda(\mathbf{F}-\mathbf{Y}))+\alpha\mathrm{tr}(\mathbf{F}^{\mathbf{T}}\Delta\mathbf{F})-\beta\mathrm{tr}(\mathbf{F}\mathbf{C}\mathbf{F}^{\mathbf{T}}), \tag{3}$$

where α and β are nonnegative constants which balance $E_s(\mathbf{F})$ and $E_c(\mathbf{F})$. For the sake of simplicity, we name this algorithm instance as ML-GRF.

We can also specify $E_l(\mathbf{F})$ and $E_s(\mathbf{F})$ as those adopted in LGC [3] method. Another graph-based algorithm can be obtained as

$$\min \mathrm{tr}((\mathbf{F} - \mathbf{Y})^{\mathbf{T}}(\mathbf{F} - \mathbf{Y})) + \mu\mathrm{tr}(\mathbf{F}^{\mathbf{T}}\mathbf{L}\mathbf{F}) - \nu\mathrm{tr}(\mathbf{F}\mathbf{C}\mathbf{F}^{\mathbf{T}}) \tag{4}$$

where $\mathbf{L}$ is the normalized combination Laplacian $\mathbf{D}^{-1/2}\Delta\mathbf{D}^{-1/2}$. μ and ν are nonnegative trade-off parameters. By minimizing Eq. (3) or solving Eq. (4), we can obtain the soft labels for unlabeled data. Next we will give the solution of ML-GRF and ML-LGC.

4.3. *Solution*

We firstly describe the solution of ML-GRF. Differentiating $E(\mathbf{F})$ with respect to $\mathbf{F}$, we have

$$\mathbf{\Lambda}(\mathbf{F} - \mathbf{Y}) + \alpha\Delta\mathbf{F} - \beta\mathbf{F}\mathbf{C} = 0. \tag{5}$$

Let $\mathbf{F} = [\mathbf{F}_l^T\mathbf{F}_u^T]^T$, and represent the matrix $\mathbf{W}$ (and similarly $\mathbf{D}$) in the form of block matrices:

$$\mathbf{W} = \begin{bmatrix} \mathbf{W}_{ll} & \mathbf{W}_{lu} \\ \mathbf{W}_{ul} & \mathbf{W}_{uu} \end{bmatrix}. \tag{6}$$

Then Eq. (5) can be written as

$$\mathbf{P}_{uu}\mathbf{F}_u + \mathbf{F}_u\mathbf{C} = \mathbf{S}, \tag{7}$$

where $\mathbf{P}_{uu} = -\alpha/\beta(\mathbf{D}_{uu} - \mathbf{W}_{uu})$ and $\mathbf{S} = -\alpha/\beta\mathbf{W}_{ul}\mathbf{Y}_l$. Eq. (7) is essentially a Sylvester equation [16], which is widely used in control theory. It is well known that Eq. (7) has a unique solution if and only if the eigenvalues $\alpha_1, \alpha_2, \cdots, \alpha_{N-L}$ of $\mathbf{P}_{uu}$ and $\beta_1, \beta_2, \cdots, \beta_k$ of $\mathbf{C}$ satisfy $\alpha_i + \beta_j \neq 0$ $(i = 1, 2, \cdots, N-L); j = (1, 2, \cdots, K)$ [16]. This condition can be easily satisfied in the real-world multi-label learning scenario.

Vectorizing the unknown matrix $\mathbf{F}_u$, Eq. (7) can be transformed to a linear equation:

$$(\mathbf{I_p} \otimes \mathbf{P}_{uu} + \mathbf{C}^T \otimes \mathbf{I_c})\mathrm{vec}(\mathbf{F}_u) = \mathrm{vec}(\mathbf{S}), \tag{8}$$

where $\otimes$ is the Kronecker product, $\mathbf{I_p}$ and $\mathbf{I_c}$ are $K \times K$ and $(N-L) \times (N-L)$ identity matrices, representatively. $\mathrm{vec}(\mathbf{M})$ is the vectorization of the matrix $\mathbf{M}$. We can then obtain $\mathbf{F}_u$ from $\mathrm{vec}(\mathbf{F}_u)$, which is equal to $(\mathbf{I_p} \otimes \mathbf{P}_{uu} + \mathbf{C}^T \otimes \mathbf{I_c})^+ \mathrm{vec}(\mathbf{S})$.

Similarly, for ML-LGC, we can obtain Eq. (9) by differentiating $E(\mathbf{F})$ with respect to $\mathbf{F}$.

$$(\mu\mathbf{L} + \mathbf{I})\mathbf{F} - \nu\mathbf{F}\mathbf{C} = \mathbf{Y}. \tag{9}$$

This equation is also a Sylvester equation and the solution is similar to that of ML-GRF.

5. Experiments

In this section, we evaluate the proposed framework on a widely used benchmark video data set and compared it against two state-of-the-art graph-based methods and one existing semi-supervised multi-label learning method.

5.1. *Implementation Issues*

In many real-world applications, the labeled data is usually insufficient. Thus the correlation matrix $\mathbf{C}$ obtained from that limited data is unreliable. To tackle this difficulty, we propose an iterative solution. In each iteration t, the labels are predicted by solving Eq. (7) (or Eq. (9)) with $\mathbf{C}^{t-1}$ estimated at last iteration $(t-1)$. Specifically, we resort to an iterative Krylov-subspace approach [30] to solve the Sylvester Equation. Then, the data points labeled with high certainty are incorporated to re-estimated $\mathbf{C}^t$.

Suppose there are M data points with the predicted label matrix $\mathbf{F}_p = [\mathbf{f}_1, \mathbf{f}_2, \cdots, \mathbf{f}_M]^T$. For the jth label, we calculate its average confidence score over all the M samples as $m_j = \frac{1}{M}\sum_{i=1}^{M} f_{ij}$, which can be token as the pseudo classification boundary of label j. The data points, which are far away from the pseudo boundary, are regarded as the samples labeled with high certainty. Specifically, for the label vector $\mathbf{f}_i$ associated with $\mathbf{x}_i$, we calculate the certainty as $g(\mathbf{f}_i) = \frac{1}{K}\sum_{j=1}^{K} \exp\{-f_{ij}(2m_j - f_{ij})\}$. The data points with large g are incorporated to re-estimate $\mathbf{C}^t$.

5.2. *Data Set Description*

To evaluate the proposed graph-based framework, we conduct the experiments for subshot-level concept detection on the benchmark TRECVID 2006 development corpus [15]. This data set contains about 170 hours international broadcast news videos from 13 international TV programs in Arabic, English and Chinese. These news videos are automatically segmented into 61,901 subshots. For each subshot, 39 concepts are multi-labeled according to LSCOM-Lite annotations [17]. These concepts were carefully chose such that they cover a variety of target types, including program category, setting/scene/site, people, object, activity, event, and graphics.

5.3. *Performance Metric*

For performance evaluation, we use the TRECVID performance metric: Average Precision (AP) [15] to evaluate and compare the approaches on each concept. Through averaging the AP over all 39 concepts, we obtain the mean average precision (MAP), which is the overall evaluation result. Specifically, Average Precision (AP) is proportional to the area under a recall-precision curve. To calculate AP for a certain concept, the test data is firstly ranked according to the prediction of each sample. Let N denote the sample number of the test set and P be the number of positive sample. At the index j, let P_j denote the number of positive subshots in the top j subshots, and I_j the indicator which indicates whether the $j^{\mathtt{th}}$ subshot is positive. The Average Precision of returned k subshots is then calculated as,

$$AP@k = \frac{1}{\min(k, P)} \sum_{j=1}^{\min(k,N)} \frac{P_j}{j} \times I_j$$

5.4. *Experimental Setup*

We conduct experiments to compare the proposed approaches (i.e., ML-GRF and ML-LGC) against other two representative graph-based methods: GRF [4] and LGC [3], and the semi-supervised multi-label learning method based on *Constrained Non-negative Matrix Factorization* (CNMF) proposed by Liu *et al.* [14].

The TRECVID 2006 development corpus is separated into four partitions, which are the same as those in [18]. This strategy of data set separation is widely used for concept detection over TRECVID 2006 corpus. Specifically, 90 videos (i.e., about 70% of the whole data set) are selected as training set, 16 videos (i.e., about 10%) as validation set, 16 videos (i.e., around 10%) as fusion set, and 15 videos (i.e., around 10%) are for performance evaluation [18]. The low-level feature we use here is 225-Dimensional block-wise color moment in Lab color space, which are extracted over 5×5 fixed grids, each block is described by 9-Dimensional features. We use the Gaussian kernel function to calculate the weighted matrix $\mathbf{W}$ for all four methods. For the sake

of fair comparison, all the algorithmic parameters in all five approaches, such as the kernel bandwidth σ and the tradeoff parameters are determined through a validation process according to their performances on the validation set. The reported performance are from the best set of parameters in all the algorithms.

When implementing the graph-based methods, we need to calculate the inversion or the multiplication of the matrix $\mathbf{W}$. Since the TRECVID 2006 development corpus contains 61,901 sub-shots, we need about 15GB memory to represent the full similarity matrix. It is difficult to store this large-scale matrix and calculate its inversion. To address this issue, we simplify the graph by only connecting neighboring samples, and thus the matrix $\mathbf{W}$ is sparse. Here we adopt K-nearest-neighbor (K-NN) method, which finds the K (K = 200) nearest neighbors for each sample, to construct the sparse representation of $\mathbf{W}$.

5.5. *Experimental Result*

Table 1 shows the MAP of our two approaches, the LGC and GRF, and Table 2 provides the detailed AP of these four approaches. The following observation can be obtained:

- By exploiting the label correlations, the proposed methods outperform the approaches which treat the semantic labels separately and neglect their integration. In details, ML-LGC achieves around 7.2% improvements on MAP compared to LGC, while ML-GRF obtains about 6.5% improvements compared to GRF.
- ML-LGC outperforms LGC over 35 of all the 39 concepts. Some of the improvements are significant, such as the 52%, 49%, and 31% improvements on "office," "airplane," and "waterscape," respectively. Compared to GRF, ML-GRF obtains improvements on 32 concepts, such as "waterscape" (105%), "office" (55%), and "airplane" (40%).
- The proposed methods degrade slightly on a few concepts. The main reason is that each of these concepts has weak interactions with other concepts. As a result, the presence/absence of these concepts cannot benefit from those of the others.

In summary, compared to the existing graph-based methods, the proposed approaches consistently outperform the performance on the diverse 39 concepts.

We also compare the proposed methods against the existing semi-supervised multi-label learning method (CNMF) [14], which has been reported to outperform some other semi-supervised multi-label learning approaches [14]. In CNMF, Liu *et al.* assumed that two data points tend to have large overlap in their assigned category memberships if they share high similarity in their input patterns. They firstly calculated two sets of similarities. The one was obtained based on the input patterns of data points, and the other was from the class memberships of the data points. Then, by minimizing the difference between these two sets of similarities, CNMF can determine the assignment of class memberships to the unlabeled data. We apply CNMF over TRECVID 2006 development set and it gets a MAP of 0.322 over the evaluation sub-set. Compared to CNMF, the proposed ML-GRF and ML-LGC methods obtain improvement of about 2.2% and 7.5%, respectively.

Table 1
Comparison of MAP for the four approaches: LGC, ML-LGC, GRF, and ML-GRF.

Approach	MAP	Improvement
LGC [3]	0.307	-
ML-LGC	**0.329**	+7.2%
GRF [4]	0.325	-
ML-GRF	**0.346**	+6.5%

Here, we further analyze the effectiveness of tradeoff parameters in the proposed methods. In ML-GRF, there exist one parameter (i.e., β/α) should be tuned. Figure 2 illustrates the MAP of ML-GRF when varying the value of β/α. The larger β/α is, the predicted labels on unlabeled samples are more consistent with the inherent label correlations, while the smaller it is, the predicted labels are more smooth over the sample graph. If we set $\beta = 0$ and $\alpha = 1$, the ML-GRF method reduces to the GRF method. Figure 3 shows the MAP of ML-LGC with respect to the tradeoff parameters μ and ν. We evaluate the influence of parameter μ by fixing ν, while μ is fixed when evaluating ν. It can be found that LGC is a special case of ML-LGC method, when $\nu = 0$.

6. Conclusions

We have proposed a novel graph-based semi-supervised learning framework to address the multi-label problems, which simultaneously takes into account both the correlations among multiple labels and the label consistency over the graph. In addition to the label smoothness, the proposed framework also addresses that the multi-label assignment to each sample should be consistent with the inherent label correlations. Furthermore, two new graph-based algorithms, which are the extension of the existing graph-based methods, were developed under the proposed framework. Experiments on the benchmark TRECVID data set demonstrated that the novel framework is superior to key existing graph-based methods and semi-supervised multi-label learning approaches, in both overall performance and the consistency of performance on diverse concepts.

We will conduct some future works from the following perspectives. First, we will apply our approaches on different data sets. Next, we will generalize it for incremental learning. Furthermore, we will combine other techniques for the multi-labeling consistency into the proposed framework.

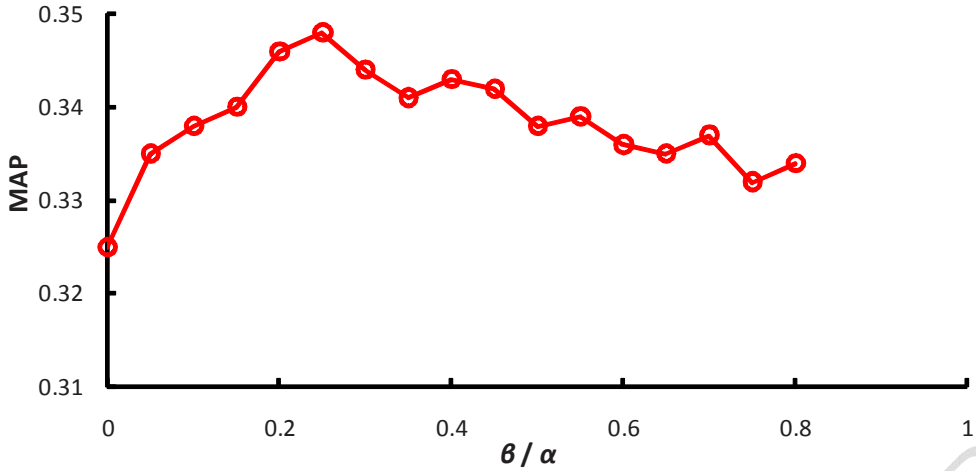

Fig. 2. Performance of ML-GRF when varying the value of the tradeoff parameter.

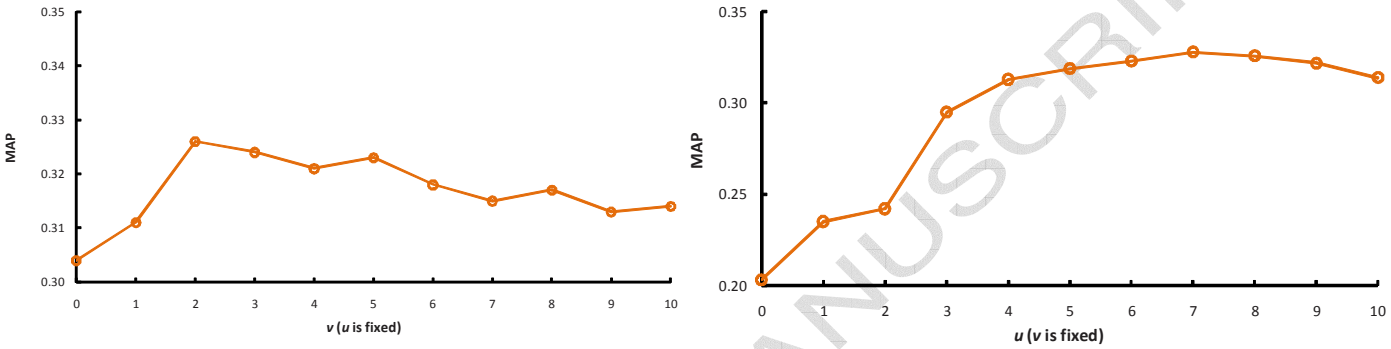

Fig. 3. Performance of ML-LGC when varying the value of the tradeoff parameters.

References

[1] G. Chen, Y. Song, F. Wang and C. Zhang, "Semi-supervised Multi-label Learning by Solving a Sylvester Equation," in *SIAM International Conference on Data Mining (SDM)*, Atlanta, Georgia, 2008, pp. 410–419.

[2] X. Zhu, "Semi-supervised learning literature survey," Tech. Rep. 1530, Computer Sciences, University of Wisconsin-Madison, 2005.

[3] D. Zhou, O. Bousquet, T.N. Lal, J. Weston, and B. Scholkopf, "Learning with local and global consistency," in *Advances in Neural Information Processing Systems (NIPS)*, Cambridge, MA, 2004, vol. 16, pp. 321–328, MIT Press.

[4] X. Zhu, Z. Ghahramani, and J. D. Lafferty, "Semi-supervised learning using gaussian fields and harmonic functions," in *Proceedings of the Twentieth International Conference on Machine Learning (ICML)*, Washington, DC, 2003, pp. 912–919.

[5] M. Belkin, I. Matveeva, and P. Niyogi, "Regularization and semi-supervised learning on large graphs," in *Annual Conference on Computational Learning Theory (COLT)*, 2004, vol. 3120 of *Lecture Notes in Computer Science*, pp. 624–638, Springer.

[6] T. Mei, X.-S. Hua, W. Lai, L. Yang, Z.-J. Zha, and *et al.*, "Msra-ustc-sjtu at trecvid 2007: High-level feature extraction and search," in *TREC Video Retrieval Evaluation Online Proceedings*, 2007.

[7] N. Ghamrawi and A. McCallum, "Collective multi-label classification," in *Proceedings of the ACM International Conference on Information and Knowledge Management (CIKM)*, New York, NY, 2005, pp. 195–200, ACM.

[8] Z.-J. Zha, T. Mei, Z. Wang and X.-S. Hua, "Building a comprehensive ontology to refine video concept detection," in *Proceedings of the international workshop on Workshop on multimedia information retrieval*, 2007, pp. 227–236, ACM.

[9] G.-J. Qi, X.-S. Hua, Y. Rui, J. Tang, T. Mei, and H.-J. Zhang, "Correlative multi-label video annotation.," in *Proceedings of the ACM International Conference on Multimedia (MM)*. 2007, pp. 17–26, ACM.

[10] J. Tang, X.-S. Hua, G.-J. Qi, M. Wang, T. Mei and X. Wu, "Structure-sensitive manifold ranking for video concept detection.," in *Proceedings of the ACM International Conference on Multimedia (MM)*. 2007, pp. 852–861, ACM.

[11] J. Tang, X.-S. Hua, G-J. Qi, Y. Song and X. Wu, "Video Annotation Based on Kernel Linear Neighborhood Propagation," in *IEEE Transaction on Multimedia*. 2008, pp. 620-628.

[12] M. Wang, T. Mei, X. Yuan, Y. Song and L.-R. Dai, "Video annotation by graph-based learning with neighborhood similarity.," in *Proceedings of the ACM International Conference on Multimedia (MM)*. 2007, pp. 325–328, ACM.

[13] Z.-J. Zha, T. Mei, J. Wang, Z. Wang and X.-S. Hua, "Graph-based semi-supervised learning with multi-label," in *Proceedings of International Conference on Multimedia and Expo (ICME)*. 2008, pp. 1321–1324.

[14] Y. Liu, R. Jin, and L. Yang, "Semi-supervised multi-label learning by constrained non-negative matrix factorization," in *Proceedings of National Conference on Artificial Intelligence and Innovative Applications of Artificial Intelligence Conference (AAAI)*, Boston, 2006, pp. 666–671.

[15] TRECVID, "http://www-nlpir.nist.gov/projects/trecvid/" .

[16] P. Lancaster and M. Tismenetsky, "The theory of matrices," *Mathematical Social Sciences*, vol. 13, no. 1, February 1987.

[17] M. Naphade, L. Kennedy, J. R. Kender, S. F. Chang, J. R. Smith, P. Over, and A. Hauptmann, "LSCOM-lite: A light scale concept ontology for multimedia understanding for TRECVID 2005," in

Table 2
Performance comparison between the proposed methods (i.e., ML-LGC and ML-GRF) and two respective graph-based approaches (i.e., LGC [3] and GRF [4]).

AP2000	LGC[3]	ML-LGC	Improve	GRF [4]	ML-GRF	Improve
Airplane	0.143	0.214	+49.7%	0.190	0.266	+40.0%
Animal	0.375	0.384	+2.4%	0.301	0.308	+2.5%
Boat_Ship	0.163	0.164	+0.6%	0.147	0.148	+0.9%
Building	0.371	0.376	+1.3%	0.310	0.338	+9.0%
Bus	0.100	0.100	-	0.093	0.095	+2.2%
Car	0.345	0.397	+15.1%	0.40	0.412	+3.0%
Charts	0.144	0.157	+9.0%	0.195	0.197	+1.2%
Computer_TV-screen	0.422	0.412	-2.4%	0.431	0.412	-4.4%
Corporate-Leader	0.016	0.017	+6.3%	0.014	0.013	-7.1%
Court	0.108	0.096	-11.1%	0.055	0.096	+74.5%
Crowd	0.271	0.29	+7.0%	0.302	0.310	+2.6%
Desert	0.134	0.146	+9.0%	0.100	0.085	-15.0%
Entertainment	0.479	0.511	+6.7%	0.512	0.531	+3.7%
Explosion_Fire	0.288	0.325	+12.8%	0.298	0.335	+12.4%
Face	0.689	0.728	+5.7%	0.751	0.758	+0.9%
Flag-US	0.042	0.056	+33.3%	0.096	0.096	-
Government-leader	0.103	0.102	-1.0%	0.144	0.137	-4.9%
Maps	0.287	0.290	+1.1%	0.454	0.456	+0.4%
Meeting	0.410	0.417	+1.7%	0.330	0.367	+11.2%
Military	0.275	0.296	+7.6%	0.287	0.304	+5.9%
Mountain	0.118	0.119	+0.8%	0.226	0.226	-
Natural-Disaster	0.410	0.424	+3.4%	0.408	0.425	+4.2%
Office	0.110	0.167	+51.8%	0.115	0.178	+54.8%
Outdoor	0.646	0.686	+6.2%	0.678	0.686	+1.2%
People-Marching	0.257	0.260	+1.2%	0.254	0.256	+0.8%
Person	0.812	0.814	+0.2%	0.816	0.817	+0.1%
Police_Security	0.008	0.010	+25.0%	0.013	0.019	+46.2%
Prisoner	0.002	0.002	-	0.002	0.002	-
Road	0.326	0.334	+2.5%	0.365	0.366	+0.3%
Sky	0.444	0.450	+1.4%	0.438	0.439	+0.2%
Snow	0.574	0.576	+0.3%	0.572	0.577	+0.9%
Sports	0.572	0.635	+11.0%	0.567	0.638	+12.5%
Studio	0.718	0.720	+0.3%	0.804	0.806	+0.2%
Truck	0.099	0.120	+21.2%	0.098	0.130	+33.3%
Urban	0.211	0.226	+7.1%	0.177	0.216	+22.0%
Vegetation	0.286	0.312	+9.1%	0.360	0.372	+3.3%
Walking_Running	0.313	0.340	+8.6%	0.295	0.333	+12.9%
Waterscape_Waterfront	0.406	0.533	+31.3%	0.240	0.491	+104.6%
Weather	0.501	0.609	+21.6%	0.849	0.850	+0.1%

Technical Report. RC23612, IBM T.J. Watson Research Center, 2005.

[18] A. Yanagawa, S.-F Chang, L. Kennedy, and W. Hsu, "Columbia university's baseline detectors for 374 LSCOM semantic visual concepts," in *Columbia University ADVENT Technical Report#222-2006-8*, 2007.

[19] A. Blum and T. Mitchell, "Combining labeled and unlabeled data with co-training," in *Proceedings of the 1998 Conference on Computational Learning Theory (COLT)*, July 1998, pp. 92–100.

[20] T. Joachims, "Transductive inference for text classification using support vector machines," in *Proceedings of the Sixteenth International Conference on Machine Learning*, San Francisco, CA, USA, 1999, pp. 200–209, Morgan Kaufmann Publishers Inc.

[21] C. Rosenberg, M. Hebert, and H. Schneiderman, "Semi-supervised self-training of object detection models," in *Seventh IEEE Workshop on Applications of Computer Vision*, January 2005, pp. 29–36.

[22] K. Nigam, A. K. McCallum, S. Thrun, and T. Mitchell, "Text classification from labeled and unlabeled documents using em," *Machine Learning*, vol. 39, no. 2-3, pp. 103–134, 2000.

[23] M. Szummer and T. Jaakkola, "Information regularization with partially labeled data," in *Advances in Neural Information Processing Systems (NIPS)*, 2002, pp. 1025–1032, MIT Press.

[24] A. K. McCallum, "Multi-label text classification with a mixture model trained by em," in *Working Notes of the AAAI'09 Workshop on Text Learning*.

[25] A. Elisseeff and J. Weston, "A kernel method for multi-labelled classification," in *Advances in Neural Information Processing Systems (NIPS)*, 2001, pp. 681–687.

[26] F. Kang, R. Jin, and R. Sukthankar, "Correlated label propagation with application to multi-label learning," in *Proceedings of IEEE International Conference on Computer*

Vision and Pattern Recognition (CVPR), 2006, pp. 1719–1726.

[27] S. Godbole and S. Sarawagi, "Discriminative methods for multi-labeled classification," in *Proceedings of Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)*, 2004, pp. 22–30.

[28] M.R. Boutell, J. Luo, X. Shen, and C.M. Brown, "Learning multi-label scene classification," *Pattern Recognition*, vol. 37, no. 9, September 2004, pp. 1757–1771.

[29] H. Kazawa, T. Izumitani, H. Taira, and E. Maeda, "Maximal margin labeling for multi-topic text categorization," in *Advances in Neural Information Processing Systems (NIPS)*, Lawrence K. Saul, Yair Weiss, and Léon Bottou, Eds., Cambridge, MA, 2005, pp. 649–656, MIT Press.

[30] D. Y. Hu and L. Reichel, "Krylov-subspace methods for the sylvester equation," *Linear Algebra and Its Applications*, vol. 172, pp. 283–313, 1992.

[31] M.-Y. Chen and A. Hauptmann, "Discriminative fields for modeling semantic concepts in video," in *Eighth Conference on Large-Scale Semantic Access to Content (RIAO)*, 2007.