Pattern Recognition 44 (2011) 2274–2286

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

A transductive multi-label learning approach for video concept detection

Jingdong Wang [a,*], Yinghai Zhao [b], Xiuqing Wu [b], Xian-Sheng Hua [a]

[a] Microsoft Research Asia, Beijing, China
[b] University of Science and Technology of China, Hefei, Anhui, China

ARTICLE INFO

Available online 13 July 2010

Keywords:
Transductive learning
Multi-label interdependence
Video concept detection

ABSTRACT

In this paper, we address two important issues in the video concept detection problem: the insufficiency of labeled videos and the multiple labeling issue. Most existing solutions merely handle the two issues separately. We propose an integrated approach to handle them together, by presenting an effective transductive multi-label classification approach that simultaneously models the labeling consistency between the visually similar videos and the multi-label interdependence for each video. We compare the performance between the proposed approach and several representative transductive and supervised multi-label classification approaches for the video concept detection task over the widely used TRECVID data set. The comparative results demonstrate the superiority of the proposed approach.

1. Introduction

Nowadays digital video content has been growing rapidly. For example, everyday a large amount of videos are uploaded to video sharing Web sites (e.g., youtube). It is very difficult for consumers to find desired videos facing with a large scale of videos, which makes video management very urgent. It is viewed as an efficient way to associate a video with semantic keywords to describe its semantic content, and then to organize videos using the associated keywords. Therefore, video annotation, a.k.a video concept detection, has emerged as a very hot research topic.

Video annotation is a process that assigns several semantic concepts to a video. Manual video annotation remains very difficult because it requires very intensive labor and too much time facing an extremely large amount of videos. Recently, human computation is viewed as a possible solution, but at present it is not mature yet to handle the annotation task of large scale videos. Consequently, more and more research focuses have been made to investigate automatic video annotation techniques.

The automatic video annotation techniques regard the annotation task as a pattern classification problem. Following the methodology of pattern classification, visual features are first extracted from a video to represent the video, and then a classifier that maps the videos to semantic keywords is learnt from a set of annotated videos. To annotate a new video, this learnt classifier is used to predict the semantic keywords. Similar to many other pattern classification problems, video annotation faces the difficulty of lacking sufficient annotated videos. Semi-supervised methods were recently developed to handle this difficulty and were also investigated in video annotation [23,28].

Unlike the traditional pattern classification problems in which each pattern has a unique label, a video may be associated with more than one label (e.g., a video may include mountain and lake.). In reality, the multiple labels appearing in a video are essentially interdependent. Considering the pairwise relationship between labels, two labels may be co-occurrent or co-absent. For example, a video with a "mountain" scene is usually also annotated as an "outdoor" concept. Some labels may be exclusive, which we call mutual–exclusive. For example, "explosion_fire" and "waterscape_waterfront" seldom occur at the same time.

Some supervised techniques [11,19,34] were developed for video annotation using the pairwise relations between labels. Several semi-supervised methods (e.g., [23,28]) discard the relations between labels, regard the labels independently, and decompose a multi-label task into a set of independent single-label tasks. A few semi-supervised approaches (e.g., [7,16,32]) tried to exploit the multi-label relations, but only exploring the multi-label inter-similarity. It leads to the label smoothness over multiple labels for each video, but discourages the mutual–exclusive relations. In this paper, we extend and analyze our proposed transductive multi-label approach [26] to handle the insufficiency of labeled videos and exploit interdependence among multiple labels.

1.1. Related work

There are many methods about semi-supervised and multi-label classification. In the following, we mainly review the closely related methods including semi-supervised single-label classification, semi-supervised multi-label classification, supervised multi-label classification and structured output classification.

* Corresponding author.
E-mail addresses: welleast@gmail.com, jingdw@microsoft.com (J. Wang).

0031-3203/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2010.07.015

Supervised methods: It is possible to extend some supervised multi-label classification methods, e.g., [11,19,30,34], to semi-supervised versions. For example, we can extend the state-of-the-art approach, correlative multi-label support vector machines (CML-SVM) [19], through some schemes similar with the single-label classification such as transductive SVM [12], semi-supervised SVM [3] and Laplacian SVM [2]. However, these extensions are computationally intractable in the large-scale case, which is also pointed out in [2]. For example, the solution for the similar extension to transductive SVM [12] requires a lot of costly iterations as each iteration needs to solve a time-consuming CML-SVM over all the data points. The method for the Laplacian extension similar to [2] suffers from solving a dense linear system which is computationally intensive. From this sense, those semi-supervised extensions are usually impractical, and hence it is necessary to pursue an effective semi-supervised multi-label approach.

Semi-supervised methods: A detailed literature survey [38,37] about semi-supervised single-label classification can be referred to for a comprehensive review. Here, we mainly discuss its application in multimedia. The classic semi-supervised methods, for example, Gaussian random field (GRF) method [39] and the local and global consistency (LGC) method [35], could be used for pattern classification directly. A learning method with unlabeled data has been used for image retrieval [36]. The co-training method has been applied to video annotation through separating the visual features carefully [20]. Furthermore, the drawbacks of the co-training based video annotation method are analyzed in [29], and an improved co-training style algorithm, called semi-supervised cross-feature learning, is proposed.

Besides, graph-based transductive classification methods have been widely applied to image and video annotation. A semi-supervised classification method based on kernel density estimation is proposed in [28] for video semantic detection, which is closely related to the graph-based approaches. In addition, several works have been conducted on introducing additional information into the graph-based method, such as the temporal consistency between the videos [22], the structure cue over the graph [23], and the multi-modality factor [27].

However, less investigation is made to address the multi-label case. There are only a few methods addressing it, e.g., [7,16,32]. However, those methods merely explore the multi-label inter-similarity, which leads to impose the smoothness assumption of the multiple labels for each data point and thus discourages the mutual–exclusive relations and hence is not accordant to the practice. Moreover, their formulations lead to a semi-definite program or a Sylvester equation, for which the computational cost is very expensive.

Structured output classification methods: The semi-supervised structured output classification problem, which is related to multi-label classification, has been studied recently in the machine learning community [1,5,9,15,40]. For example, in [1], a maximum margin semi-supervised learning approach is proposed in the standard reproducing kernel Hilbert space (RKHS) framework. It imposes the smoothness assumption over the labels, which intuitively means to only bias the co-positive or co-negative relations. In [15,40], the semi-supervised approaches to structured variables are presented in the RKHS framework and kernel conditional random field framework, respectively. A semi-supervised approach is proposed for structured input and output or multi-classification in [5], but only exploits multi-view information to minimize the disagreement for unlabeled examples without exploring the interdependence of the structured output. The above methods are related to multi-label classification, but are very different as the input is also a structured variable with each element corresponding to an element of the output, while the input of multi-label classification is a single variable. Therefore, the above methods cannot be directly applied to the multi-label classification problems, such as the video concept detection problem in the TRECVID tasks.

1.2. Contributions

In this paper, we present a discrete hidden Markov random field approach for transductive multi-label classification. Our approach aims to find a labeling, such that it satisfies two properties similar to the existing transductive single-label classification methods: (1) the labeling is the same as the pre-given labeling over the labeled data points; (2) the labeling is consistent between the data points with similar features; and particularly, an extra property: (3) the multi-label interdependence over the unlabeled data points is coherent with that over the labeled data points, and consists of not only the co-positive and co-negative relations, but also the mutual–exclusive relations. In this paper, the models for the three aspects are presented in Sections 4.2–4.4, respectively.

We illustrate the difference with the conventional graph-based transductive single-label classification approach in Fig. 1. Compared with the possible semi-supervised extension of

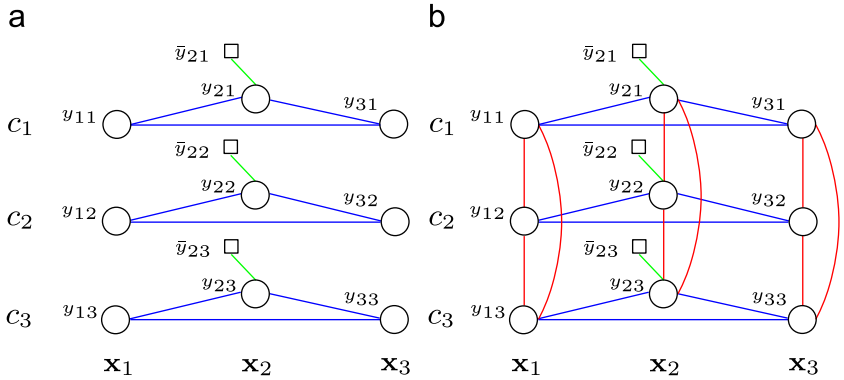

Fig. 1. Illustration of the difference between the conventional transductive approach and our approach for multi-label classification in the form of hidden Markov random fields. ○ is a hidden variable, which represents a binary-valued label corresponding to the associated sample and concept. □ is an observable variable, which is a pre-given binary-valued label over the labeled data point. The conventional approach is based on several independent graphs as shown in (a). Our approach considers the multi-label interdependence which is depicted in (b). The edge connecting a pre-given label node and its associated hidden label node, e.g., $(y_{21},\bar{y}_{21})$, depicts the compatibility between the pre-given and predicted labeling over the labeled data point. Here, we assume $\mathbf{x}_2$ is a labeled data point. The edge connecting two nodes over hidden labels corresponding to the same concept, e.g., (y_{11}, y_{21}), represents the labeling consistency between the neighboring data points. Particularly, the edge connecting two nodes over hidden labels corresponding to the same sample, e.g., (y_{31}, y_{33}) is introduced by the proposed approach to model the multi-label interdependence.

supervised multi-label classification methods, the proposed approach directly infers the labeling through an efficient optimization algorithm without the necessity to estimate the intermediate inductive decision function. In summary, this paper mainly offers the following contributions:

- A hidden Markov random field framework is proposed for the transductive multi-label classification problem, which simultaneously models the local labeling consistency and multi-label interdependence and comes with an efficient labeling inference.
- The multi-label interdependence is modeled by a pairwise Markov random field. All the combinations of pairwise relations over the multiple labels are seamlessly modeled in a single integrated formulation, including the positive–positive, positive–negative, negative–positive, and negative–negative relations.
- We analyze the connection of our approach with several existing graph-based transductive single-label classification approaches and the Sylvester equation based transductive multi-label classification approach, and conclude that they can essentially be viewed as specified instances of our framework.

2. Problem setup

In this paper, we represent one video by the visual feature $\mathbf{x}$ extracted from it. In the semi-supervised multi-label setting, we are given n videos, $\mathcal{X}=\{\mathbf{x}_i\}_{i\in\mathcal{N}}$, $\mathcal{N}=\{1,\ldots,n\}$, and l videos, $\{\mathbf{x}_i\}_{i\in\mathcal{L}}$, are assigned with K-dimensional binary-valued label vectors $\{\overline{\mathbf{y}}_i\}_{i\in\mathcal{L}}$, $\overline{\mathbf{y}}_i\in\{0,1\}^K$, $\mathcal{L}=\{1,\ldots,l\}$. Here, K is the cardinality of the label set $\mathcal{C}=\{c_1,\ldots,c_K\}$. $\overline{y}_{ic}=1$ indicates that $\mathbf{x}_i$ is associated with the concept c, and is not associated with it otherwise. The task is to assign label vectors $\{\mathbf{y}_i\}_{i\in\mathcal{U}}$ to the remaining videos, $\{\mathbf{x}_i\}_{i\in\mathcal{U}}$, $\mathcal{U}=\{l+1,\ldots,n\}$. Denote $u=n-l$ as the number of unlabeled videos, $\mathbf{Y}=[\mathbf{y}_1\ \ldots\ \mathbf{y}_n]^T$ a label matrix of size $n\times k$, and $\mathbf{y}^c=\mathbf{Y}(:,c)$ represents the c-th column vector and corresponds to a labeling configuration with respect to the concept c. $\overline{\mathbf{Y}}_l$ and $\mathbf{Y}_u$ correspond to the labeling over the labeled and unlabeled data, respectively. It should be noted that there may be more than 1 entries valued by 1 for $\mathbf{y}_i$ in the multi-label case, while there is only one entry valued by 1 in the multi-class case.

3. Transductive single-label classification

In this section, we review the transductive single-label classification methods. Semi-supervised classification methods were proposed to leverage both the labeled and unlabeled data. There are many semi-supervised classification methods, and a detailed survey of the literature on semi-supervised classification methods is presented in [37,38]. In general, these methods can be divided into two categories: inductive classification methods and transductive classification methods. Inductive classification methods try to induce a decision function that has a low classification error rate on the whole sample space, whereas transductive classification methods directly estimate the labels of the unlabeled data.

In the following, we briefly review graph-based transductive methods, which are closely related to our proposed approach. The common assumption of graph-based methods is label consistency [35,25] (also called cluster assumption [6]), i.e., nearby points are prone to have the same labels. Most existing graph-based methods essentially estimate a labeling on a graph that is constructed by connecting the neighboring data points, expecting that such a labeling satisfies two properties: (1) It should be as close as possible to (or the same as) the given labeling on the labeled data points, and (2) it should be smooth on the whole graph. The former property can be formulated as a loss function penalizing the deviation of the predicted labeling from the given labeling. The latter property can be expressed as a regularizer enforcing label consistency.

A typical graph-based semi-supervised classification method is illustrated in Fig. 1(a). In the graph, the nodes consist of the label nodes, including the labels of both labeled and unlabeled videos, and the nodes, each of which corresponds to the label node and shows the prior of the value of the label node. The edges consist of a set of edges, each of which connects two label nodes and is weighted by the similarity between the corresponding two videos to impose the labeling consistency of the two videos, and another set of edges, each of which connects a label node and the corresponding pre-given label node.

4. Transductive multi-label classification

In this section, we will present the proposed transductive multi-label classification approach. As mentioned before, the transductive single-label classification method aims to find a labeling that satisfies two properties: (1) It should be as close as possible to (or the same as) the given labeling on the labeled data points, and (2) it should be smooth on the whole graph. In the multi-label case, an additional information, multi-label dependence, is helpful. Hence, the labeling in multi-label case should satisfy an extra property that the multi-label interdependence on the unlabeled data points should be similar to that on the labeled data points.

We present a discrete hidden Markov random field (dHMRF) formulation, which integrates all the three properties into a single framework and makes it natural to introduce and model the multi-label interdependence. We use the graph terminology to describe the dHMRF construction. Next, we reformulate the loss function and the regularization function in the dHMRF framework. Then, we present the key part, multi-label dependence. Finally, we aggregate them together to get the overall formulation.

4.1. Graph construction

A hidden node is constructed for each entry y_{ic}, and denoted by y_{ic} for simplicity. The set of hidden nodes is denoted as $\mathcal{V}_h$. An observable node is associated with each labeled data point, and denoted as $\overline{y}_{ic}$. The set of observable nodes is denoted as $\mathcal{V}_o$. The observable nodes and their corresponding hidden nodes are connected to construct a set of edges, called the observable edges and denoted as $\mathcal{E}_o=\{(y_{ic},\overline{y}_{ic})\}_{i\in\mathcal{L},c\in\mathcal{C}}$.

In addition, we build a sample-pair-wise edge by connecting y_{ic} and y_{jc} for $c\in\mathcal{C}$ if their corresponding data points $\mathbf{x}_i$ and $\mathbf{x}_j$ are neighboring points. In this paper, the neighboring points are determined by k-nearest neighbor graph, which leads to at least two advantages: local smoothness and low computation cost. We denote this edge set as $\mathcal{E}_s=\{(y_{ic},y_{jc})\}_{i,j\in\mathcal{L}\cup\mathcal{U},c\in\mathcal{C}}$. Particularly, we build an edge set on all the node pairs, $(y_{i\alpha},y_{i\beta})$. Intuitively, we connect the pair of nodes corresponding to the same data point. We denote the edge set as $\mathcal{E}_d=\{(y_{i\alpha},y_{i\beta})\}_{i\in\mathcal{U},\alpha,\beta\in\mathcal{C}}$, and call it label-pair-wise edges.

The constructed graph is shown in Fig. 1(b). Strictly speaking, the points $\{\mathbf{x}_i\}$ should also be involved into the graph structure as the dependent nodes of the hidden variables $\{y_{ic}\}$, but to make the presentation convenient and easily understood, we omit them in the formulation and mention it only where we essentially use the data points $\{\mathbf{x}_i\}$. In the following, we mathematically define the potential functions over all the edges on the dHMRF.

4.2. Loss function

We first consider the consistency between the final labeling and the pre-labeling over the given videos, and define a set of loss functions separately on the observable edges $\{(y_{ic},\overline{y}_{ic})\}$, as shown in Fig. 1(b). In this paper, we only give a penalization when the final labeling is different from the given labeling, and define the loss function in the following form:

$$e_{ic}(y_{ic}) = \gamma_{ic}\mathbf{1}[y_{ic} \neq \overline{y}_{ic}], \tag{1}$$

where γ_{ic} is used to set a penalty if y_{ic} is not equal to $\overline{y}_{ic}$, and $\mathbf{1}[\cdot]$ is an indicator function. We aggregate all the functions over all the observable edges and get an overall loss function,

$$\mathrm{E}_{\mathrm{loss}} = \sum_{i \in \mathcal{L}, c \in \mathcal{C}} e_{ic}(y_{ic}) \tag{2}$$

$$= \sum_{i \in \mathcal{L}, c \in \mathcal{C}} \gamma_{i,c}\mathbf{1}[y_{ic} \neq \overline{y}_{ic}]. \tag{3}$$

The loss function can be transformed to a probability formulation based on a *Boltzmann* distribution of an energy function,

$$\mathrm{O}(y_{ic}) = \frac{\exp(-e_{ic}(y_{ic}))}{\exp(-e_{ic}(0)) + \exp(-e_{ic}(1))}. \tag{4}$$

In this paper, we assume that the pre-labeling is correct, equivalently viewing the pre-labeling as hard constraints. Specifically, in the loss function, we give an infinity penalty if y_{ic} is not equal to $\overline{y}_{ic}$, i.e., $\gamma_{ic} = \infty$. This is equivalent to a discrete probability over the latent variable y_{ic},

$$\mathrm{O}(y_{ic}) = \mathbf{1}[y_{ic} = \overline{y}_{ic}]. \tag{5}$$

Then, we obtain an overall probability based on the *Boltzmann* distribution of the overall loss function E_{loss}

$$\mathrm{O}(\mathbf{Y}_l) \propto \exp(-E_{\mathrm{loss}}) \tag{6}$$

$$\propto \prod_{i \in \mathcal{L}, c \in \mathcal{C}} \mathbf{1}[y_{ic} = \overline{y}_{ic}]. \tag{7}$$

4.3. Local labeling regularization

The local labeling regularization term aims to penalize the case that the labeling is not locally consistent. In other words, it is expected that the videos with similar visual features have the same labeling. In our problem, a video may be associated with more than one concepts, i.e., a label vector with binary valued entries. We define the labeling consistency (smoothness) between multiple labels, by factorizing the smoothness over label vectors and aggregating the smoothness value of each label together.

Specifically, we define the regularization function as an Ising energy over an interacting pair y_{ic} and y_{jc},

$$e_{ij,c}(y_{ic},y_{jc}) = \gamma^0_{ij,c}\mathbf{1}[y_{ic} \neq y_{jc}] + \gamma^1_{ij,c}\mathbf{1}[y_{ic} = y_{jc}], \tag{8}$$

where $\gamma^0_{ij,c}$ is a penalty that only one of $\mathbf{x}_i$ and $\mathbf{x}_j$ is assigned the label c, and $\gamma^1_{ij,c}$ is a penalty that both $\mathbf{x}_i$ and $\mathbf{x}_j$ are assigned or not assigned the label c. Then the overall regularization function over all the videos and the associated concepts is defined as

$$E_{\mathrm{reg}} = \sum_{(i,j) \in \mathcal{E}^c_s, c \in \mathcal{C}} e_{ij,c}(y_{ic},y_{jc}). \tag{9}$$

Similarly to the last subsection, the regularization function can be transformed to an equivalent probability function,

$$r(y_{ic},y_{jc}) = \sum_{p,q \in \{0,1\}} P^{pq}_{ij,c}\mathbf{1}[y_{ic} = p \wedge y_{jc} = q], \tag{10}$$

where $\sum_{p,q \in \{0,1\}} P^{pq}_{ij,c} = 1$, $P^{pq}_{ij,c}$ is the probability when $y_{ic} = p$ and $y_{jc} = q$ and can be calculated from the *Boltzmann* distribution of Eq. (8). A probability function over the label vector is written as

$$r(\mathbf{y}_i,\mathbf{y}_j) = \prod_{c \in \mathcal{C}} r(y_{ic},y_{jc}). \tag{11}$$

Combining all the potentials over the sample-pair-wise edges, we obtain the aggregated probability,

$$R(\mathbf{Y}) \propto \prod_{(i,j) \in \mathcal{E}^c_s, c \in \mathcal{C}} r(y_{ic},y_{jc}). \tag{12}$$

Here, it should be pointed out that the formulation in the consistency probability $R(\mathbf{Y})$ is simplified and strictly it should be written as $R(\mathbf{Y};\{\gamma^0_{ij,c},\gamma^1_{ij,c}\})$ and the parameters $\{\gamma^0_{ij,c},\gamma^1_{ij,c}\}$ are dependent on $\{\mathbf{x}\}_i$.

In this paper, similar to [39], we define the two penalties (γ) from two aspects: (1) it is preferable that the labeling is as smooth as possible, (2) the penalty of the non-consistency between two points is proportional to the similarity. Therefore, we let $\gamma^1_{ij,c} = 0$, $\gamma^0_{ij,c} = w_{ij}$, $w_{ij} = \exp(-\theta\|\mathbf{x}_i - \mathbf{x}_j\|^2)$ with $\theta > 0$ being a kernel parameter and the Gaussian function is selected as the graph of the Gaussian is a characteristic symmetric "bell shape curve" that quickly falls off towards plus/minus infinity (here, the distance $\|\mathbf{x}_i - \mathbf{x}_j\|$ between $\mathbf{x}_i$ and $\mathbf{x}_j$).

4.4. Multiple label grouping and interdependence

Some existing semi-supervised methods, such as in [23,27], have been directly applied to K-label classification. The basic scheme is factorizing it into K independent semi-supervised binary classification problems and inferring the labels for the K concepts, respectively. From the perspective of probability theory, it factorizes the joint distribution $\Pr(\mathbf{Y})$ into the product of K independent probabilities, which is equivalent to defining a probability over the K isolated subgraphs, $\{\mathcal{G}_1,\ldots,\mathcal{G}_K\}$, formed only by the observable and sample-pair-wise edge sets, with each corresponding to a concept. Obviously, this method neglects the interdependence over the multiple labels.

In real tasks, e.g., the TRECVID data set, we observed that some concepts have close dependence or little dependence. For example, in the TRECVID data set, "bus" and "car" have strong mutual-dependence, but they have less dependence with "weather". This motives us to divide the concepts into several categories such that the concepts from different categories, called concept chunklets, have smaller interdependence and the concepts in the same category have larger interdependence. The categorization helps to decouple some noisy relations between concepts, which may be due to the data set limitation and leads to robust estimation of the dependence between multi-concepts. On the other hand, it is very helpful to fast inference. In our experiments, the overall inference is about 80 times faster than the supervised multi-label approach in [19].

In the following, we present a two-step scheme to model the dependence among multiple labels into our formulation. We first categorize the concepts into a few groups according to the relations between the concepts. Then, we present the interdependence formulation for each group of concepts.

4.4.1. Concept grouping

First, we compute the relation degree between each pair of concepts using the normalized mutual information (NMI) measure. Specifically, we select the geometric mean based normalized symmetric mutual information because this measure is analogous to a normalized inner product in the Hilbert space [21]. In summary, such an NMI has the following properties. (1) It is symmetric. (2) Its range is [0, 1], and it attains a minimum value of zero when the variables are independent, and a maximum value of 1, for example when the variables are completely decided

mutually; And (3) it is adoptable for our concept grouping algorithm, spectral clustering, which usually uses a radial basis function (RBF). RBF's range is also [0, 1], and it is similar to the used NMI. Such an NMI is defined as

$$\mathrm{NMI}(\alpha,\beta)=\frac{I(\alpha,\beta)}{\sqrt{H(\alpha)H(\beta)}}. \tag{13}$$

Here $I(\alpha,\beta)$ is the mutual information between the two labels, α and β, and defined as

$$I(\alpha,\beta)=\sum_{\alpha,\beta\in\{0,1\}}\Pr(\alpha,\beta)\log\frac{\Pr(\alpha,\beta)}{\Pr(\alpha)\Pr(\beta)}, \tag{14}$$

and $H(\alpha)$ is the entropy of α and defined as

$$H(\alpha)=-\sum_{\alpha\in\{0,1\}}\Pr(\alpha)\log\Pr(\alpha). \tag{15}$$

The joint probability, $\Pr(\alpha,\beta)$, is evaluated as

$$P_{\alpha\beta}^{pq}=\frac{1}{|\mathcal{L}|}\sum_{i\in\mathcal{L}}\mathbf{1}[y_{i\alpha}=p\wedge y_{i\beta}=q], \tag{16}$$

where $p,q\in\{0,1\}$, and $\mathbf{1}[\cdot]$ is an indicator function. In practice, to avoid the too small probability due to the insufficiency of the labeled data points, we correct the probabilities as the following, $P_{\alpha\beta}^{pq}=P_{\alpha\beta}^{pq}+\varepsilon$, where ε is a small constant that is valued by $100/n$ in our experiment, and normalize them such that the summation is equal to 1. The marginal probability, $\Pr(\alpha)$, is similarly as

$$P_{\alpha}^{p}=\frac{1}{|\mathcal{L}|}\sum_{i\in\mathcal{L}}\mathbf{1}[y_{i\alpha}=p]. \tag{17}$$

To group the concepts, we adopt the spectral clustering method [18]. We build an affinity matrix, $\mathbf{A}=[a_{ij}]$, with $a_{ij}=\mathrm{NMI}(i,j)$. To get a well-separated concept chunklets, we use the normalized Laplacian matrix, $\overline{\mathbf{A}}=\mathbf{D}^{1/2}(\mathbf{D}-\mathbf{A})\mathbf{D}^{1/2}$, where $\mathbf{D}=\mathrm{Diag}([d_1,\ldots,d_n]^T)$ and $d_i=\sum_j a_{ij}$. We eigen-decompose $\overline{\mathbf{A}}$, so that $\overline{\mathbf{A}}=\mathbf{V}^T\mathbf{\Lambda}\mathbf{V}$, and then use the spectral analysis technique in [18] to investigate the eigenvalues for the estimation of the cluster number m. Finally, we represent each concept by a m-dimensional vector $\mathbf{z}$, with $\mathbf{z}$ formulated by the second to $m+1$ entries of the corresponding row vector $\mathbf{V}$, and run k-means over these representative vectors to obtain the concept clusters, which are denoted by a set of concept groups, $\{\mathcal{C}_k\}_{k=1}^{N}$.

In our experiment, we found that a few concepts that have large relation degrees with all the other concepts, which we call common concepts. Before grouping the concepts, we first discover these common concepts from the relation degrees. The relation degree for each concept is calculated as $d_\alpha=\sum_{c\in\mathcal{C}}\mathrm{NMI}(\alpha,c)$. According to the degrees as shown in Fig. 2, we can see that the first three degrees are much larger than the others'. Hence, their corresponding concepts, "face", "outdoor" and "person", are selected as the common concepts. The concept grouping result over the remaining concepts is depicted in Fig. 3.

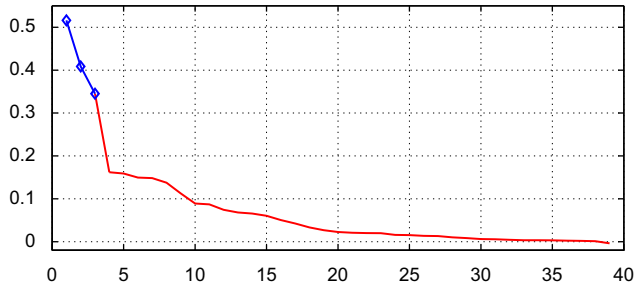

Fig. 2. Illustration of the ordered relation degrees of all the 39 concepts. It can observed that the first three degrees are much larger than the others. This implies that the corresponding three concepts have strong relations with other concepts.

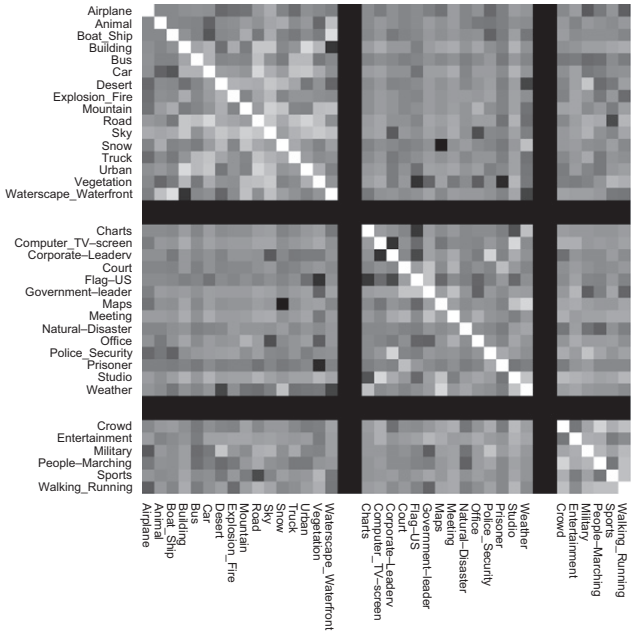

Fig. 3. The normalized mutual information (NMI) between all the pairs over the 36 concepts are depicted. Three chunklets are separated by black bars. It can be observed that the NMI between the labels in the same chunklet, seated on the diagonal blocks, is larger and that the NMI between the labels from different chunklets, seated on the off-diagonal blocks, is much smaller.

4.4.2. Label interdependence

We have grouped the multiple concepts into several groups $\{\mathcal{C}_k\}_{k=1}^{N}$. Accordingly, we partition the label vector into several subvectors, $\{\mathbf{y}_k\}_{k=1}^{N}$. We define the interdependence between multiple labels for each group of concepts.

We formulate the pairwise interdependence between the multiple labels as a pairwise Markov random field. Specifically, we define the interdependence probability as a product of the potential functions over all the pairs of labels for each data point,

$$D_k(\mathbf{y}_{k,i})=\frac{1}{Z'}\prod_{\alpha,\beta\in\mathcal{C}_k}\phi(y_{i\alpha},y_{i\beta}), \tag{18}$$

where $\phi(y_{i\alpha},y_{i\beta})$ is the interactive probability of the membership of the data point i with respect to two concepts α and β,

$$\phi(y_{i\alpha},y_{i\beta})=\sum_{p,q\in\{0,1\}}P_{\alpha\beta}^{pq}\mathbf{1}[y_{i\alpha}=p\wedge y_{i\beta}=q]. \tag{19}$$

In this sense, the label-pair-wise edges essentially bridge all the isolated subgraphs associated with the concepts and aggregate them into a single graph.

Different from the potential function estimation in the sample-pair-wise edges, we learn the functions over the label-pair-wise edges from the labeled data points in the maximum likelihood criterion, i.e., maximizing

$$\prod_{i\in\mathcal{L}}D_k(\mathbf{y}_{k,i})\propto\prod_{i\in\mathcal{L},\alpha,\beta\in\mathcal{C}_k}\phi(y_{i\alpha},y_{i\beta}). \tag{20}$$

The estimation is NP-hard since our MRF is a loopy graph. In [14], an approximate iterative algorithm is presented for estimation. In this paper, we present a fast and effective estimation scheme by using the joint probability over a pair of labels to estimate the potential function,

$$P_{\alpha\beta}^{pq}=\frac{1}{|\mathcal{L}|}\sum_{i\in\mathcal{L}}\mathbf{1}[y_{i\alpha}=p\wedge y_{i\beta}=q], \tag{21}$$

where $p,q \in \{0,1\}$, and $\mathbf{1}[\cdot]$ is an indicator function. This estimation is faster than the method in [14], but its performance is almost the same with the performance from [14]. In practice, to avoid the too small probability due to the insufficiency of the labeled data points, we correct the probabilities as the following, $P_{\alpha\beta}^{pq} = P_{\alpha\beta}^{pq} + \varepsilon$, where ε is a small constant that is valued by $100/n$ in our experiment, and normalize them such that the summation is equal to 1. Then, all the potentials over the label-pair-wise edges for the unlabeled data are joined together as

$$D_k(\mathbf{Y}_{k,u}) = \prod_{i \in \mathcal{U}} \mathrm{D}_k(\mathbf{y}_{k,i}) \tag{22}$$

$$\propto \prod_{i \in \mathcal{U}, \alpha, \beta \in \mathcal{C}_k} \phi(y_{i\alpha}, y_{i\beta}). \tag{23}$$

By integrating several groups of concepts together, we get the overall probability,

$$D(\mathbf{Y}_u) = \prod_k \mathrm{D}_k(\mathbf{Y}_{k,u}) \tag{24}$$

$$\propto \prod_k \prod_{i \in \mathcal{U}, \alpha, \beta \in \mathcal{C}_k} \phi(y_{i\alpha}, y_{i\beta}). \tag{25}$$

It is worth to point out that the multi-label interdependence that we take into accounts consists of all possible relations between the concept pairs. In [32], only the multi-label inter-similarities are taken into accounts, which leads to a bias of the co-positive and co-negative relations between the concept pairs. This is not accordant to the practice, e.g., in the TRECVID data set, since the mutual–exclusive relations occur frequently. For example, the concepts "airplane" and "sky" often co-exist while "explosion_file" and "waterscape_waterfront" seldom occur at the same time. In our approach, we explore all the relations among the labels in such a way that the co-positive and co-negative relations, the mutual–exclusive relations, including the negative–positive and positive–negative relations, are reasonably captured.

4.5. Overall objective

The above three probabilities are aggregated into a single probability,

$$\begin{aligned}\Pr(\mathbf{Y}) &= \frac{1}{Z}\mathrm{O}(\mathbf{Y}_l)\mathrm{R}^{\lambda}(\mathbf{Y})\mathrm{D}^{1-\lambda}(\mathbf{Y}_u) \\ &= \frac{1}{Z}\prod_{i \in \mathcal{L}, c \in \mathcal{C}} \mathrm{O}(y_{ic}) \prod_{c \in \mathcal{C}, (i,j) \in \mathcal{E}_s^c} r^{\lambda}(y_{ic}, y_{jc}) \prod_k \prod_{i \in \mathcal{U}, \alpha, \beta \in \mathcal{C}_k} \phi(y_{i\alpha}, y_{i\beta}),\end{aligned} \tag{26}$$

where λ is a trade-off variable to control the balance between the sample-pair-wise and label-pair-wise probabilities, and Z is a normalization constant. The corresponding dHMRF is illustrated in Fig. 1(b). Here, the first term of the right-hand side corresponds to the compatibility between the hidden labels and the pre-labeling, i.e., loss function. The second term corresponds to the label consistency between the data points with similar features, i.e., local labeling regularization. The last term corresponds to the multi-label interdependence. The solution to the proposed transductive multi-label classification is found as the joint maximum,

$$\mathbf{Y}^* = \arg\max_{\mathbf{Y}} \Pr(\mathbf{Y}). \tag{27}$$

It should be pointed out that in the formulation of $\Pr(\mathbf{Y})$ the conditional variables, $\{\mathbf{x}\}_i$ and $\{\overline{y}_{ic}\}$ and the parameters, $\{\gamma_{ij,c}^0, \gamma_{ij,c}^1\}$ and $\{P_{\alpha\beta}^{pq}\}$ are omitted for presentation convenience.

4.6. Inference

The optimization for the discrete hidden Markov random field in Eq. (26) is nontrivial. Many optimization methods have been developed, such as the augmenting directed acyclic graph algorithm, simulate annealing, iterated conditional modes, loopy belief propagation, tree-reweighted message passing, graph cuts and so on. The detailed discussion on those algorithms can be found in [10].

In this paper, we combine tree-reweighted message passing [13] and graph cuts [4] to infer the labels efficiently. The first one, which is an iterative statistical inference algorithm, guarantees that a near-optimal solution is obtained. This algorithm iteratively propagates vector valued messages along the edges connecting the nodes until convergence and finally can find the maximum *a posteriori* solution. The speed of the tree-reweighted message passing algorithm depends on the initialization. Therefore, the second one is adopted to provide a good initialization, a *warm start*, to speed up the inference. Graph cuts is an algorithm to separate a graph by finding a cut (a set of edges) of the graph such that the weight on the edges is minimum, and implemented by finding a maximum flow of the graph from the source to the sink. It is guaranteed to obtain the global optimal solution when the pairwise energy function is submodular. For example, in Eq. (8), $e_{ij,c}(0,0)+e_{ij,c}(1,1) \leq e_{ij,c}(0,1)+e_{ij,c}(1,0)$ holds, which means $e_{ij,c}$ is submodular, and non-submodular otherwise. In our case, the label-pair-wise energy is not guaranteed to satisfy this submodular property. Therefore, we discard the non-submodular energies and perform the graph cuts algorithm to obtain a solution, which is used as the warm start.

The output of the proposed transductive multi-label classification method is binary-valued. For the video retrieval evaluation, an ordered score is often required. With these scores, the retrieved videos can be ranked. We present a probabilistic scoring scheme from the discrete output. The score is evaluated as a conditional probability. Given the solution $\mathbf{Y}^*$, the score of $y_{ic} = 1$, i.e., the data point $\mathbf{x}_i$ is associated with the label c, is evaluated as

$$\begin{aligned}&\Pr(y_{ic} = 1 | \mathbf{Y}_{\setminus y_{ic}} = \mathbf{Y}^*_{\setminus y_{ic}}) \\ &= \Pr(y_{ic} = 1 | y_{jc} = y^*_{jc}, (i,j) \in \mathcal{E}_s^c, y_{i\beta} = y^*_{i\beta}, \beta \in \mathcal{C} \setminus \{c\}) \\ &= \frac{1}{Z} \prod_{(i,j) \in \mathcal{E}_s^c} r^{\lambda}(y_{ic} = 1, y_{jc} = y^*_{jc}) \prod_{c \in \mathcal{C}_k, \beta \in \mathcal{C}_k \setminus \{c\}} \phi^{1-\lambda}(y_{ic} = 1, y_{i\beta} = y^*_{i\beta}),\end{aligned} \tag{28}$$

where Z is a normalization constant, and $\mathbf{Y}_{\setminus y_{ic}}$ is all the entries in $\mathbf{Y}$ except the entry y_{ic}.

5. Connections

In this section, we analyze the connections between our approach and several related methods, and conclude that the Gaussian random field (GRF) approach in [39] and the Sylvester equation based approach in [32] are reduced versions of our approach. Moreover, we discuss the difference of our approach from correlative multi-label support vector machines (CML-SVM) [19].

5.1. Connection with GRF

In this subsection, we re-discuss the local labeling consistency formulation presented in Section 4.3 and show that GRF is essentially a relaxed version of our formulation. To make the description clearer, we only consider the regularization $E_{\mathrm{reg},\,c}$ over a single label c in Eq. (9), and obtain the following formulation by dropping the subscript c:

$$E_{\mathrm{reg}} = \sum_{ij} e_{ij}(y_i, y_j). \tag{29}$$

Considering $y_i \in \{0,1\}$ and $\gamma_{ij}^1 = 0$ in Eq. (8), we transform it to an equivalent formulation

$$e_{ij}(y_i,y_j) = \gamma_{ij}^0 \mathbf{1}[y_i \neq y_j] = \gamma_{ij}^0 (y_i - y_j)^2. \quad (30)$$

Consequently, the regularization is transformed to the following discrete function:

$$E_{\text{reg}} = \sum_{ij} \gamma_{ij}^0 (y_i - y_j)^2. \quad (31)$$

Suppose $\gamma_{ij}^0 = w_{ij}$ and discard the integer constraint $y_i \in \{0,1\}$, we can obtain the GRF formulation

$$E_{\text{reg}} = \sum_{ij} w_{ij} (y_i - y_j)^2. \quad (32)$$

From this sense, GRF is a degradation of our approach by discarding the multi-label interdependence and relaxing discrete labels to continuous labels. If we define asymmetric probabilities over the sample-pair-wise edges, the local and global consistency (LGC) method in [35] can also be viewed as a reduced version of our approach.

5.2. Connection with the Sylvester equation

In this subsection, we show that the proposed approach is a generalized version of the Sylvester equation, which is used in [32,33] to solve the transductive multi-label problem. First, we transform the multi-label interdependence probability in Eq. (19) to the equivalent energy formulation,

$$e_{i,\alpha\beta}(y_{i\alpha},y_{i\beta}) = \sum_{p,q \in \{0,1\}} \gamma_{\alpha\beta}^{pq} \mathbf{1}[y_{i\alpha} = p \wedge y_{i\beta} = q], \quad (33)$$

where $\gamma_{\alpha\beta}^{pq} = -\log(P_{\alpha\beta}^{pq})$. We replace the energies by the label similarity $w_{\alpha\beta}$: $\gamma_{\alpha\beta}^{00} = \gamma_{\alpha\beta}^{11} = 0$ and $\gamma_{\alpha\beta}^{01} = \gamma_{\alpha\beta}^{10} = w_{\alpha\beta}$. The two energies are depicted in Table 1. Considering $y_{i\alpha} \in \{0,1\}$, the above energy function is equivalently transferred to

$$e_{i,\alpha\beta}(y_{i\alpha},y_{i\beta}) = w_{\alpha\beta} \mathbf{1}[y_{i\alpha} \neq y_{i\beta}] = w_{\alpha\beta} (y_{i\alpha} - y_{i\beta})^2. \quad (34)$$

Then the energy function for the multi-label interdependence over a point have the following matrix formulation:

$$\mathbf{y}_i^T (\mathbf{D}_d - \mathbf{W}_d) \mathbf{y}_i, \quad (35)$$

where $\mathbf{W}_d = [w_{\alpha\beta}]_{K\times K}$, and $\mathbf{D}_d = \text{Diag}(d_1,\ldots,d_K)$ with $d_c = \sum_{\alpha \in C} w_{c\alpha}$. Considering all the points, the energy function is written as

$$E_{\text{dep}} = \text{trace}(\mathbf{Y}\mathbf{\Delta}_d\mathbf{Y}^T), \quad (36)$$

where $\mathbf{\Delta}_d = \mathbf{D}_d - \mathbf{W}_d$. The regularization term for the local label consistency has a matrix form

$$E_{\text{reg}} = \text{trace}(\mathbf{Y}^T\mathbf{\Delta}_s\mathbf{Y}), \quad (37)$$

Table 1
Comparison of the energies on the concept-pair-wise edges. (a) Corresponds to our approach, (b) corresponds to the Sylvester equation based approach.

$e_{i,\alpha\beta}$	$y_{i\alpha} = 0$	$y_{i\alpha} = 1$
(a)		
$y_{i\beta} = 0$	$\gamma_{\alpha\beta}^{00}$	$\gamma_{\alpha\beta}^{01}$
$y_{i\beta} = 1$	$\gamma_{\alpha\beta}^{10}$	$\gamma_{\alpha\beta}^{11}$
(b)		
$y_{i\beta} = 0$	0	$w_{\alpha\beta}$
$y_{i\beta} = 1$	$w_{\alpha\beta}$	0

Note that the preference to different label relations in our approach is learnt from the training data, while the Sylvester equation based approach prefers only the co-positive and co-negative relations between the concepts, which is unreasonable and inconsistent to the practice.

where $\mathbf{\Delta}_s$ is the graph Laplacian operator over the sample-pair-wise edges. Relaxing discrete labels into continuous values and differentiating $E_{\text{reg}} + E_{\text{dep}}$ with respect to $\mathbf{Y}$, we have

$$\lambda\mathbf{\Delta}_s\mathbf{Y} + (1-\lambda)\mathbf{Y}\mathbf{\Delta}_d = 0, \quad (38)$$

and discard the given labels, we have a Sylvester equation,

$$\lambda\mathbf{A}\mathbf{Y}_u + (1-\lambda)\mathbf{Y}_u\mathbf{B} = -\lambda\mathbf{C}\overline{\mathbf{Y}}_l, \quad (39)$$

where $\mathbf{Y}_u$ and $\overline{\mathbf{Y}}_l$ are the label matrices corresponding to the unlabeled data points and labeled data points. $\mathbf{A}$ is a $u \times u$ submatrix of $\mathbf{\Delta}_s$ corresponding to the unlabeled data points, $\mathbf{B} = \mathbf{\Delta}_d$, and $\mathbf{C}$ is a $u \times l$ submatrix of $\mathbf{\Delta}_s$.

In summary, the Sylvester equation based approach in [32,33] is also a specified version of our approach, which adopts the label similarity that leads to an unreasonable preference to the co-positive and co-negative relations, and relaxes the discrete labels to continuous numbers, but does not capture the mutual–exclusive relations between the concepts.

5.3. Difference from CML-SVM

This subsection discusses the difference between our approach and the correlative multi-label support vector machine approach (CML-SVM), proposed in [19]. The common point is that both the approaches utilize the multi-label interdependence to help the classification. However, they are different in the following aspects. First, CML-SVM is a purely supervised approach, which cannot leverage the unlabeled data as analyzed in Section 1. Differently, our approach is a semi-supervised approach, which essentially aims to deal with the insufficiency of the labeled data. Second, the learning scheme of multi-label interdependence is different. Our approach offers a generative way rather than a discriminative way. Finally, our approach is more efficient due to the fast combinatorial optimization approach while CML-SVM is quite computationally expensive. Hence, our approach can be applied in real problems, particularly in large-scale problems.

6. Experiments

6.1. Setup

We evaluate the proposed transductive multi-label classification approach on the development set of the TRECVID 2005 high-level feature extraction task. This set contains 137 broadcast news videos with 43,907 shots (further segmented into 61,901 sub-shots) from 13 different programs in English, Arabic and Chinese [24]. Following the data set separation scheme presented in [31], we divide this data set into four partitions: the training set (67.6%), the validation set (11.34%), the fusion set (10.54%), and the testing set (10.52%). The fusion set is not suitable for our experiments, and hence is discarded. Therefore, we have 41,847 sub-shots for training (labeled data), 7022 sub-shots for validation, and 6507 sub-shots for testing (unlabeled data). For each sub-shot, we extract a low-level feature from the corresponding key-frame. Some example key frames are shown in Fig. 5. In our experiments, we use a 225-dimensional block-wise color moment, which is shown to be able to obtain better performance on the TRECVID data set than the local features (e.g., scale invariant feature transform (SIFT) descriptors). This feature is extracted over 5×5 blocks with each block described by a

9-dimensional feature, and transform each dimension such that it satisfies a standard normal distribution.

According to the LSCOM-Lite annotations [17], 39 concepts are multi-labeled for each sub-shot. These annotated concepts consist of a wide range of genres, including program category, setting/scene/site, people, object, activity, event, and graphics. Many of these concepts have significant semantic dependence between each other, which can be observed from Fig. 3. As pointed in [19], many sub-shots (71.32%) have more than one label, and some sub-shots are even labeled with as many as 11 concepts. The two observations motivate us to explore the multi-label interdependence in the transductive classification problem.

For performance evaluation, we adopt the widely used performance metric, average precision (AP), in the TRECVID task. We calculate the AP value for each concept according to their ranking score. Then we average the APs over all the 39 concepts to create the mean average precision (MAP) to evaluate the overall performance of a specific classifier.

6.2. Results

In our experiments, we perform three types of comparative experiments with three kinds of methods:

1. The state-of-the-art of supervised single-label and multi-label classification method.
2. The popular graph-based transductive single-label classification method.
3. One representative transductive multi-label classification method.

In the following experiments, for our approach, abbreviated by TML, the graph structure of the hidden Markov random field is constructed as follows. The sample-pair-wise edges are constructed by connecting a video and its 30 nearest neighbors. The three common concepts, "face", "outdoor" and "person", are

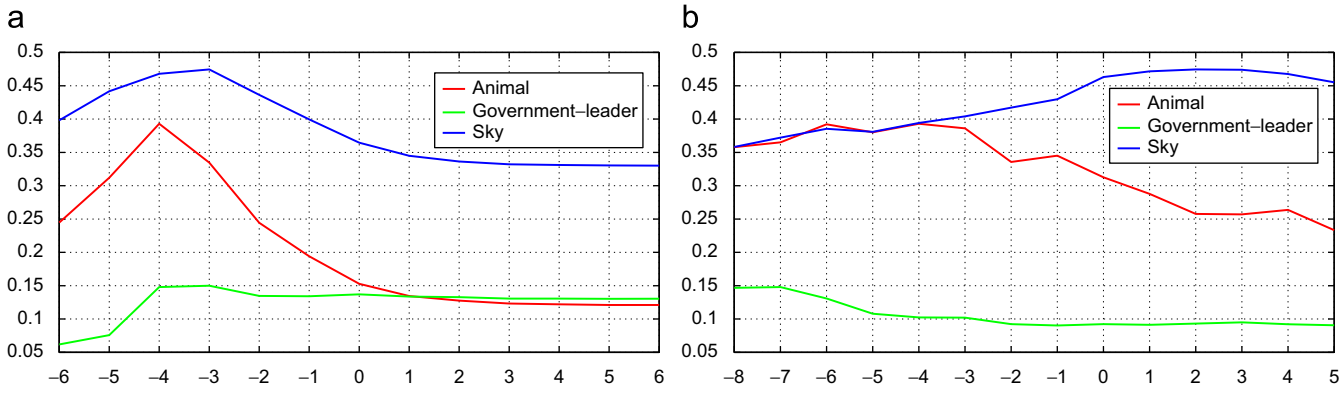

Fig. 4. Illustration of the performance variance with different kernel parameters and trade-off parameters. In both figures, the y-axis represents the average precision. (a) shows the performance variance with tuning the kernel parameter θ, and the x-axis indicates the base 2 logarithm of θ. (b) shows the performance variance with tuning the trade-off parameter λ, and the x-axis indicates the base 2 logarithm of λ.

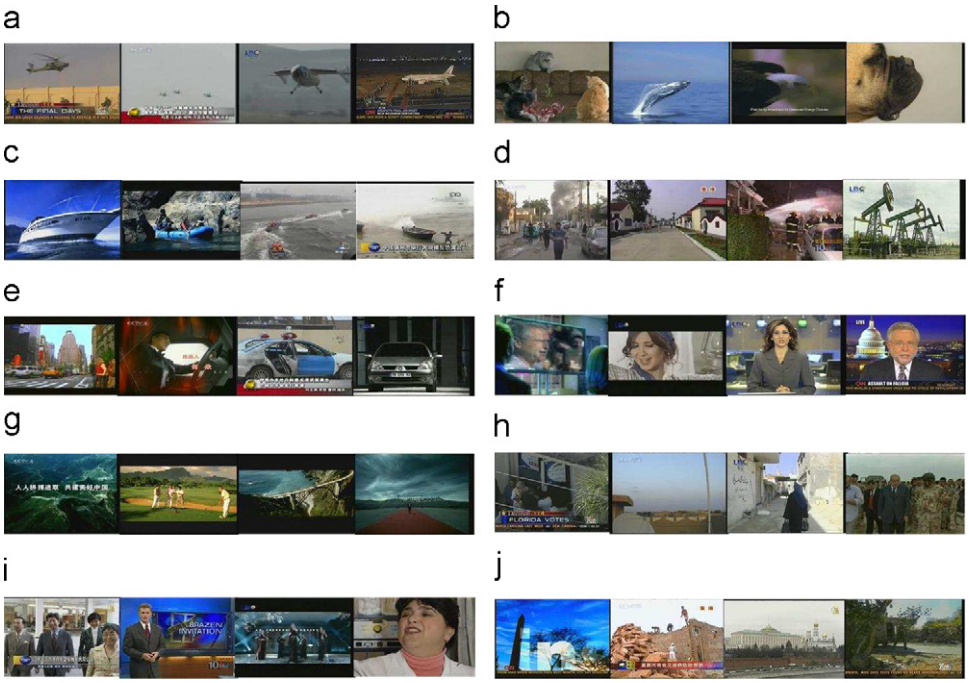

Fig. 5. Exemplary key frames for several randomly sampled concepts. (a) airplane, (b) animal, (c) boat ship, (d) building, (e) car, (f) face, (g) mountain, (h) outdoor, (i) person and (j) sky.

added to all the three concept chunklets forming three augmented chunklets. The concepts in each augmented chunklet are connected with each other to build the label-pair-wise edges. Therefore, three subgraphs are obtained. Intuitively, only the interdependence between the labels from the same chunklets is taken into account, and the dependence between the labels from different chunklets is very minor and hence neglected. Then we run our approach over the three graphs respectively. The APs of the three common concepts are taken as the maximum APs in the three augmented chunklets.

For implementation, the kernel parameter θ and the trade-off variable λ between the sample-pair-wise and label-pair-wise probabilities are required to be pre-determined. The performance variance with different parameters is depicted in Fig. 4. It can be seen that different parameters lead to very different performances. Therefore, in our experiment, the parameters are determined through a validation process. We select the parameters corresponding to the best performances on the validation set. Then we report the performances over the test set with those selected parameters. The top 10 results are shown in Fig. 6. In this figure, the false-positive key frames are indicated with bounding boxes.

6.2.1. *Vs. supervised single-label and multi-label classification*

In this experiment, we compare the proposed approach with the state-of-the-art of supervised single-label classification, support vector machines (SVM), and supervised multi-label classification proposed in [19], correlative multi-label classification (CML) with the data setup being the same as that of our approach. For SVM, we adopt the Gaussian kernel and the parameters are obtained by a validation process. For CML, we adopt the Gaussian kernel used in [19]. The Gaussian kernel parameter and the trade-off variable in CML are obtained through the validation process, which is the same with [19].

The comparative results are presented in Table 2. We can see that our approach outperforms CML. This superiority mainly comes from the transductive scheme, which directly estimates the labeling based on the local labeling consistency without

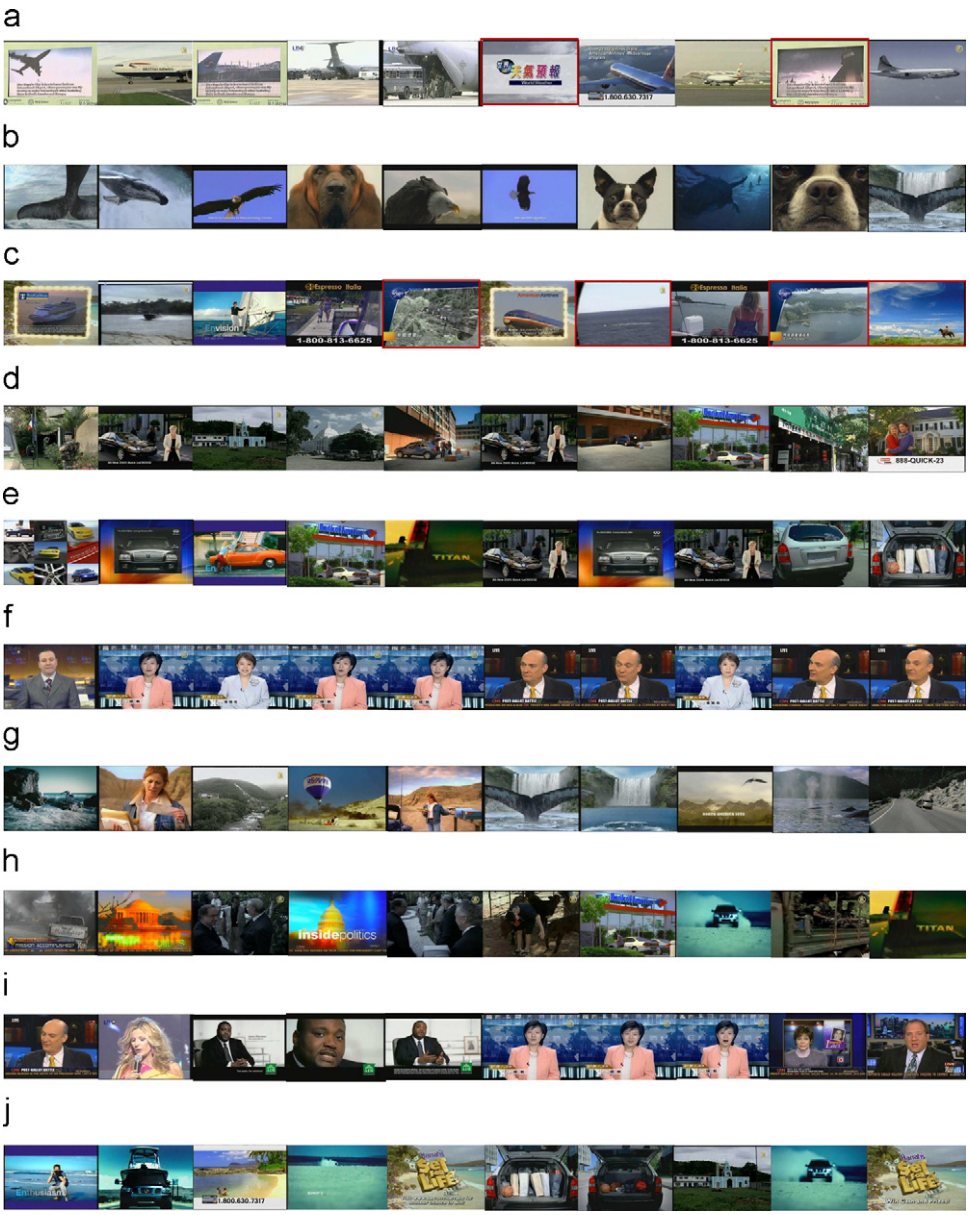

Fig. 6. Visual results. Key frames of the top 10 sub-shots for some randomly sampled concepts. (a) airplane, (b) animal, (c) boat ship, (d) building, (e) car, (f) face, (g) mountain, (h) outdoor, (i) person and (j) sky.

estimating an intermediate classifier. Thus the following observations can be obtained:

- TML obtains 8.81% and 5.1% relative improvements over SVM and CML.
- TML outperforms SVM and CML on 21 and 26 of 39 concepts. For some concepts, TML obtains significant improvements. For example, compared with SVM, the improvements over "boat_ship", "charts", "court", "desert" and "government_leader", are 568.7%, 169.5%, 3518.3%, 284.3% and 196.3%.
  Compared with CML, the relative improvements on "airplane", "corporate-leader", "mountain", and "truck" are 112%, 127%, 116%, and 116%, respectively.
- TML gets lower performances on some concepts, such as "face", "person" "maps", "office", and "studio", than CML. This deterioration is due to the local labeling consistency constraint, which smoothes some too high AP scores.

6.2.2. Vs. transductive single-label classification

In this experiment, we compare the proposed approach with two popular graph-based transductive single-label classification approaches, including the Gaussian random field (GRF) method [39], the local and global consistency (LGC) method [35], and the structure-sensitive manifold ranking (SSMR) method [23].

The graph for GRF, LGC and SSMR is constructed just by removing all the label-pair-wise edges in the dHMRF graph, and hence consists of 39 subgraphs with each corresponding to a concept. In GRF and LGC, the weights on the sample-pair-wise edges are defined as a Gaussian kernel that is similar to our approach. In SSMR, the density of the point is estimated by the Parzen window method. The kernel parameter for the three methods is selected through a similar validation process.

In addition, we conduct an experiment on the 39 subgraphs using the similar objective function to GRF but keeping the label integer constraints, which is equivalent to our objective function with removing the multi-label interdependence. Then, we perform the optimization (only graph cuts is performed as the energy is submodular in this case) and the presented scoring scheme. In the experiment, we denote this method by TSL.

The comparative results are presented in Table 2, from which it can be observed that multi-label interdependence is capable of

Table 2
Performance comparison.

Concepts	Semi-supervised single-label				Supervised multi-label		Transductive multi-label	
	LGC	GRF	SSMR	TSL	SVM	CML	Sylvester	TML
Airplane	0.142961	0.1598	0.212745	0.1788	**0.359693**	0.110213	0.1705	0.2334
Animal	0.374792	0.3496	0.351143	0.3001	0.353494	0.288026	0.3222	**0.3929**
Boat_Ship	0.163279	0.1191	0.147397	**0.1757**	0.0206079	0.109042	0.1365	0.1378
Building	**0.371038**	0.2604	0.272293	0.2969	0.294832	0.278689	0.3332	0.2819
Bus	0.096047	0.06619	0.092572	0.08192	**0.0997387**	0.0670332	0.08146	0.09336
Car	0.345209	0.3989	0.390521	0.3796	0.420418	**0.432184**	0.3806	0.3906
Charts	0.143641	0.146	0.194371	0.1893	0.0730492	0.15287	0.1882	**0.1969**
Computer_TV-screen	0.422452	0.4184	0.436511	0.4313	0.440921	0.433267	0.365	**0.4449**
Corporate_Leader	0.016225	**0.06992**	0.010096	0.01205	0.0592525	0.0121081	0.01115	0.02752
Court	**0.108213**	0.07608	0.055157	0.0502	0.00213459	0.0758106	0.05176	0.07725
Crowd	0.270527	0.3099	0.301387	0.2531	0.392024	**0.403026**	0.2728	0.3394
Desert	0.134344	0.08343	0.133281	0.1447	0.0449079	0.113796	0.1515	**0.1726**
Entertainment	0.479208	0.526	**0.53945**	0.4453	0.429239	0.506431	0.4922	0.5009
Explosion_Fire	0.287927	0.2916	**0.304092**	0.3005	0.209007	0.213329	0.2674	0.2657
Face	0.688516	0.7605	0.771883	0.7441	0.756963	**0.80511**	0.7428	0.75
Flag_US	0.041525	**0.1418**	0.097926	0.04633	0.0974388	0.12549	0.05502	0.1322
Government_leader	0.103064	0.1454	0.146079	0.1232	0.0499186	**0.200731**	0.09732	0.1479
Maps	0.287281	0.4807	0.44132	0.5245	**0.622993**	0.577103	0.4744	0.529
Meeting	0.41046	0.2338	0.391134	0.2305	0.199684	**0.429509**	0.2305	0.3616
Military	0.275304	0.2829	0.287284	0.2766	**0.347068**	0.265552	0.3201	0.3415
Mountain	0.11775	0.1113	0.231373	0.2386	0.100874	0.11376	0.1637	**0.2454**
Natural_Disaster	0.409932	0.317	0.410229	0.3763	**0.532321**	0.390069	0.4032	0.402
Office	0.110148	0.1661	0.117937	0.1168	**0.282566**	0.162967	0.09543	0.1145
Outdoor	0.645513	0.6854	0.683848	0.6298	0.629509	**0.695713**	0.6746	0.6811
People_Marching	0.257356	0.2518	0.256847	**0.2976**	0.203306	0.175272	0.2634	0.2887
Person	0.812266	0.8204	0.832895	0.7797	**0.861693**	0.840236	0.8041	0.7962
Police_Security	0.007649	0.01101	**0.014238**	0.01258	0.00842659	0.00886924	0.01133	0.01087
Prisoner	0.001817	0.0033	0.002121	0.00136	0.000375115	0.00144856	0.00126	**0.00357**
Road	0.325824	**0.3833**	0.371385	0.3743	0.332742	0.38082	0.3749	0.373
Sky	0.444037	0.4221	0.437512	0.4188	**0.536747**	0.479817	0.4509	0.4744
Snow	0.573949	0.4664	0.577571	0.5564	**0.601854**	0.485786	0.5353	0.5889
Sports	0.572139	0.5783	0.549632	0.5871	0.623732	0.614891	0.6145	**0.6434**
Studio	0.717921	0.8324	0.817487	0.8015	0.827177	**0.847251**	0.8103	0.8216
Truck	0.099039	0.08316	0.08185	0.1397	0.15953	0.0942839	0.08546	**0.2036**
Urban	0.211396	0.1963	0.181537	0.2093	**0.222713**	0.190943	0.214	0.212
Vegetation	0.286339	0.3528	0.350812	0.3622	**0.404011**	0.342246	0.3456	0.3539
Walking_Running	0.313089	0.3003	0.31536	**0.3176**	0.112857	0.259197	0.2976	0.3159
Waterscape_Waterfront	0.40573	0.3017	0.443875	0.466	0.398521	0.481982	0.455	**0.4833**
Weather	0.501122	**0.8718**	0.852665	0.8652	0.472158	0.85798	0.8412	0.8631
mAP	0.307	0.32	0.336	0.327	0.322679408	0.333919	0.32273	**0.351**
Improvements	14.33	9.69	4.49	7.34	8.81	5.09	8.57	–

The average precision (AP) over each concept, the mean AP (mAP) over all the 39 concepts, and the relative mAP improvement of our approach TML over other approaches are presented.

improving the concept detection performance. The detailed analysis is given as the following:

- Our approach (TML) gets an MAP of 0.351 and obtains about 14.3%, 9.7% and 4.49% relative improvements, compared with LGC, GRF and SSMR. TSL obtains a similar performance with GRF. This shows that the major factor of the performance improvement of our approach lies in the multi-label interdependence.
- TML performs better on 31, 26 and 29 of all the 39 concepts than LGC, GRF and SSMR. The improvements of some concepts are significant. For instance, the improvements on "airplane", "mountain", and "truck" are 63%, 108%, and 106% compared with LGC, and 46%, 120%, 145% compared with GRF. For the concepts whose SSMR APs are smaller than 0.3, TML obtains steadily improvements on the annotation performance. For example, the improvements on "airplane", "crowd", "desert" and "truck", compared with SSMR, are 9.7%, 12.6%, 29.5% and 148.7%, respectively.
- TML has a slightly deteriorated performance on the common concepts, e.g., "face", "outdoor", "person", which have strong relations with almost all other concepts. Compared with the best performance of single-label methods, there are 2.91% deterioration on "face", 0.6% deterioration on "outdoor", 7.69% deterioration on "entertainment", 14.4% "explosion_fire", and 4.6% deterioration on "person". The deterioration is because the high performances over these concepts are pulled down by the relatively lower performances of the other concepts.

6.2.3. Vs. the discrete Sylvester equation

In this experiment, we compare our approach with the Sylvester equation based transductive multi-label classification approach (denoted by Sylvester) in [32]. This method always prefers to the co-positive or co-negative relations for any pair of labels, which is not consistent to the practice and fails to capture the mutual–exclusive relations between the concepts. Moreover, solving a Sylvester equation is computationally expensive when the scale of the data set is large. For fairness, we do not relax the discrete labels to continuous numbers but keep it to a discrete optimization problem, which can be solved by the algorithm presented in this paper. The whole implementation is the same with ours, except the multi-label interdependence. The comparison is presented in Table 2. The detailed comparison is as follows:

- TML gains an 8.8% relative improvement over Sylvester.
- TML improves the performances on 32 concepts, and outperforms Sylvester with overwhelmingly superiority over some concepts. For example, TML outperforms Sylvester over "corporate-leader", "flag-US", and "truck" by 147%, 140% and 138%, respectively.
- There is a little deterioration over a few concepts. For example, the performances on "building", "person", and "police-security", are deteriorated by 1.8%, 1%, and 4%, respectively.

The MAPs of the above existing approaches and the improvements of our approach over the other methods are summarized in Table 2. It can be seen that our approach get the best overall performance and the improvements are notable.

In addition, we would like to present the comparison on the computational cost. Counting both training and predicting processes, among all the approaches, the time cost of TSL is the lowest, and the costs of our approach, TML, and Sylvester are similar and a little higher than TSL. This is understandable as the three approaches both adopt discrete inference algorithms and the graph structure of TSL is simpler than those of TML and Sylvester. The costs of LGC, GRF, and SSMR are larger because the discrete inference (adopted by TML and Sylvester), in practice, is more efficient than the numerical inference (adopted by LGC, GRF, SSMR). The costs of SVM and CML are the highest as the training process has to solve a time-consuming quadratic program. Then, let us consider the computational cost without considering the training process, which can be performed offline. The cost of individual SVM is the lowest. This is reasonable because other approaches have to involve the labeled data or consider multi-label relations, which makes the prediction more time-consuming. Finally, we would like to give a detailed comparison between CML and TML. For the training and predicting process, the whole time cost of our approach is much less (about one-twentieth) than CML, because the training process of CML is very costly while our approach has no such process and directly predicts the labels for unlabeled data. Without considering the training process, the time costs of the two approaches are comparable because both our approach and the expensive step of CML need to solve combinatorial optimization problems, and CML even involves an inductive step. In detailed analysis, our approach solves a combinatorial optimization problem over all the unlabeled data and their neighboring labeled data, and CML solves a series of combinatorial optimization problem over all individual unlabeled data points. For our problem, we present an efficient inference method that consists of a fast graph-cuts algorithm and an efficient tree-reweighted belief propagation

Table 3
Illustration of out-of-sample extension.

Concepts	Transductive	Out of sample
Airplane	0.09881	**0.09916**
Animal	0.4076	**0.4084**
Boat_Ship	0.5813	**0.5816**
Building	**0.2338**	0.2334
Bus	**0.009816**	0.009788
Car	0.2671	**0.268**
Charts	**0.1886**	0.1885
Computer_TV-screen	0.4354	**0.4385**
Corporate-Leader	0.1019	**0.1024**
Court	0.2264	**0.3201**
Crowd	0.3815	**0.3926**
Desert	**0.1262**	0.1261
Entertainment	0.326	**0.3274**
Explosion_Fire	0.08845	0.08845
Face	0.7916	**0.796**
Flag-US	0.02756	**0.2189**
Government-leader	0.1881	**0.2343**
Maps	0.491	**0.5302**
Meeting	0.167	**0.1816**
Military	0.3841	**0.3896**
Mountain	**0.2161**	0.216
Natural-Disaster	0.11	0.11
Office	**0.117**	0.1169
Outdoor	0.6771	**0.6883**
People-Marching	0.3645	**0.367**
Person	**0.8204**	0.8187
Police_Security	0.1028	**0.1078**
Prisoner	0.006864	**0.008608**
Road	0.2217	**0.2243**
Sky	0.4862	**0.492**
Snow	**0.1541**	0.129
Sports	**0.4683**	0.4681
Studio	0.821	**0.8233**
Truck	**0.1158**	0.1152
Urban	**0.2046**	0.2023
Vegetation	**0.3776**	0.3758
Walking_Running	**0.1855**	0.1822
Waterscape_Waterfront	**0.3437**	0.3281
Weather	0.8245	**0.8247**
mAP	0.3113	**0.32141**

algorithm with a warm start, but the inference method for CML is a little time-consuming due to the complexity of the problem.

6.3. Out-of-sample extension

In this subsection, we propose a method to extend our approach to the out-of-sample data points. In the out-of-sample setting, when new data points appear, the out-of-sample extension technique can be adopted to infers the labels of the out-of-sample data points without the necessity of retraining and without re-predicting the labels of the previous unlabeled data points, thus reduce the computation cost. According to [8], we need to address two issues: (1) use the same type of regularizer for a new testing point, and (2) the inclusion of the new testing point should not affect the original labeling of the in-sample data points. We run the proposed algorithm again to solve the task of labeling the out-of-sample data points by viewing both the labeled data and unlabeled data as the new labeled data and viewing the out-of-sample data points as the unlabeled data. We test the out-of-sample extension technique by viewing the fusion set as out-of-sample data points, and the result is presented in Table 3. We also report the result in Table 3 by integrating the fusion and testing set together and running our algorithm. From the results, the out-of-sample extension technique did not reduce the performance. Furthermore, the computational cost of the out-of-sample extension technique for the fusion set is only about one-fifth of the retraining procedure. Through this extension, our approach can be adopted for large-scale problems.

7. Conclusions

This paper presents a probabilistic transductive multi-label approach to together handle the insufficiency of labeled videos and the multiple labeling problem that are two important issues in video concept detection. The proposed approach simultaneously models the labeling consistency between visually similar videos and the multi-label interdependence for each video in an integrated framework. Moreover, we present a label chunklet analysis scheme to make the labeling inference more efficient. The performance comparison with several existing approaches on the widely used TRECVID data set shows the superiority of our approach.

References

[1] Y. Altun, D.A. McAllester, M. Belkin, Maximum margin semi-supervised learning for structured variables, in: NIPS, 2005.
[2] M. Belkin, P. Niyogi, V. Sindhwani, Manifold regularization: a geometric framework for learning from labeled and unlabeled examples, Journal of Machine Learning Research 7 (2006) 2399–2434.
[3] K.P. Bennett, A. Demiriz, Semi-supervised support vector machines, in: NIPS, 1998, pp. 368–374.
[4] Y. Boykov, V. Kolmogorov, An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision, IEEE Transactions on Pattern Analysis and Machine Intelligence 26 (9) (2004) 1124–1137.
[5] U. Brefeld, T. Scheffer, Semi-supervised learning for structured output variables, in: ICML, 2006, pp. 145–152.
[6] O. Chapelle, J. Weston, B. Schölkopf, Cluster kernels for semi-supervised learning, in: NIPS, 2002, pp. 585–592.
[7] G. Chen, Y. Song, F. Wang, C. Zhang, Semi-supervised multi-label learning by solving a sylvester equation, in: SDM, 2008.
[8] O. Delalleau, Y. Bengio, N.L. Roux, Efficient non-parametric function induction in semi-supervised learning, Society for Artificial Intelligence and Statistics, 2005, pp. 96–103.
[9] K. Duh, K. Kirchhoff, Structured multi-label transductive learning, in: NIPS Workshop on Advances in Structured Learning for Text/Speech Processing, 2005.
[10] B.J. Frey, N. Jojic, A comparison of algorithms for inference and learning in probabilistic graphical models, IEEE Transactions on Pattern Analysis and Machine Intelligence 27 (9) (2005) 1392–1416.
[11] W. Jiang, S.-F. Chang, A.C. Loui, Active context-based concept fusion with partial user labels, in: ICIP, 2006, pp. 2917–2920.
[12] T. Joachims, Transductive inference for text classification using support vector machines, in: ICML, 1999, pp. 200–209.
[13] V. Kolmogorov, Convergent tree-reweighted message passing for energy minimization, IEEE Transactions on Pattern Analysis and Machine Intelligence 28 (10) (2006) 1568–1583.
[14] H.H. Ku, S. Kullback, Approximating discrete probability distributions, IEEE Transactions on Information Theory (IT) 15 (4) (1969) 444–447.
[15] C.-H. Lee, S. Wang, F. Jiao, D. Schuurmans, R. Greiner, Learning to model spatial dependency: semi-supervised discriminative random fields, in: NIPS, 2006, pp. 793–800.
[16] Y. Liu, R. Jin, L. Yang, Semi-supervised multi-label learning by constrained non-negative matrix factorization, in: AAAI, 2006.
[17] M.R. Naphade, L. Kennedy, J.R. Kender, S.-F. Chang, J.R. Smith, P. Over, A. Hauptmann, A light scale concept ontology for multimedia understanding for TRECVID 2005, in: IBM Research Report RC23612 (W0505-104), 2005.
[18] A.Y. Ng, M.I. Jordan, Y. Weiss, On spectral clustering: analysis and an algorithm, in: NIPS, 2001, pp. 849–856.
[19] G.-J. Qi, X.-S. Hua, Y. Rui, J. Tang, T. Mei, H.-J. Zhang, Correlative multi-label video annotation, in: ACM Multimedia, 2007, pp. 17–26.
[20] Y. Song, X.-S. Hua, L.-R. Dai, M. Wang, Semi-automatic video annotation based on active learning with multiple complementary predictors, in: Multimedia Information Retrieval, 2005, pp. 97–104.
[21] A. Strehl, J. Ghosh, Cluster ensembles—a knowledge reuse framework for combining multiple partitions, Journal of Machine Learning Research 3 (2002) 583–617.
[22] J. Tang, X.-S. Hua, T. Mei, G.-J. Qi, X. Wu, Video annotation based on temporally consistent Gaussian random field, Electronics Letters 43 (8) (2007).
[23] J. Tang, X.-S. Hua, G.-J. Qi, M. Wang, T. Mei, X. Wu, Structure-sensitive manifold ranking for video concept detection, in: ACM Multimedia, 2007, pp. 852–861.
[24] TRECVID2005, 2005. 〈http://www-nlpir.nist.gov/projects/trecvid/〉.
[25] J. Wang, F. Wang, C. Zhang, H.C. Shen, Long Quan, Linear neighborhood propagation and its applications, IEEE Trans. Pattern Anal. Mach. Intell. 31 (9) (2009) 1600–1615.
[26] J. Wang, Y. Zhao, X. Wu, X.-S. Hua, Transductive multi-label learning for video concept detection, in: Multimedia Information Retrieval, 2008, pp. 298–304.
[27] M. Wang, X.-S. Hua, X. Yuan, Y. Song, L.-R. Dai, Optimizing multi-graph learning: towards a unified video annotation scheme, in: ACM Multimedia, 2007, pp. 862–871.
[28] M. Wang, Y. Song, X. Yuan, H. Zhang, X.-S. Hua, S. Li, Automatic video annotation by semi-supervised learning with kernel density estimation, in: ACM Multimedia, 2006, pp. 967–976.
[29] R. Yan, M.R. Naphade, Semi-supervised cross feature learning for semantic concept detection in videos, in: CVPR, vol. 1, 2005, pp. 657–663.
[30] R. Yan, M. yu Chen, A.G. Hauptmann, Mining relationship between video concepts using probabilistic graphical models, in: ICME, 2006, pp. 301–304.
[31] A. Yanagawa, S.-F. Chang, L. Kennedy, W. Hsu, Columbia University's baseline detectors for 374 LSCOM semantic visual concepts, Technical Report, Columbia University, March 2007.
[32] Z.-J. Zha, T. Mei, J. Wang, Z. Wang, X.-S. Hua, Graph-based semi-supervised learning with multi-label, in: ICME, 2008, pp. 1321–1324.
[33] Z.-J. Zha, T. Mei, J. Wang, Z. Wang, X.-S. Hua, Graph-based semi-supervised learning with multiple labels, Journal of Visual Communication and Image Representation 20 (2) (2009) 97–103.
[34] M.-L. Zhang, Z.-H. Zhou, ML-KNN: a lazy learning approach to multi-label learning, Pattern Recognition 40 (7) (2007) 2038–2048.
[35] D. Zhou, O. Bousquet, T.N. Lal, J. Weston, B. Schölkopf, Learning with local and global consistency, in: NIPS, 2003.
[36] Z.-H. Zhou, Learning with unlabeled data and its application to image retrieval, in: PRICAI, 2006, pp. 5–10.
[37] Z.-H. Zhou, M. Li, Semi-supervised learning by disgreement. Knowledge and Information Systems, 2010, doi:10.1007/s10115-009-0209-z.
[38] X. Zhu, Semi-supervied learning literature survey. Computer Sciences Technical Report, 1530, University of Wisconsin-Madison, 2007.
[39] X. Zhu, Z. Ghahramani, J. Lafferty, Semi-supervised learning using Gaussian fields and harmonic functions, in: ICML, 2003, pp. 912–919.
[40] A. Zien, U. Brefeld, T. Scheffer, Transductive support vector machines for structured variables, in: ICML, 2007, pp. 1183–1190.

Jingdong Wang received the B.Sc. and M.Sc. degrees in automation from Tsinghua University, Beijing, China, in 2001 and 2004, respectively, and the Ph.D. degree in computer science from the Hong Kong University of Science and Technology, Hong Kong, in 2007. He is currently an associate researcher in the Media Computing Group at Microsoft Research Asia. His areas of interest include machine learning, pattern recognition, multimedia computing, and computer vision. In particular, he has worked on kernel methods, semi-supervised learning, data clustering, image segmentation, image and video presentation, and organization and search.

Yinghai Zhao is a Ph.D. student at University of Science and Technology of China, Hefei, Anhui, China. His research interests include multimedia and pattern recognition.

Xiuqing Wu received the B.S. degree from the University of Science and Technology of China, Hefei, in 1965. Currently, she is a Professor in the Department of Electronic Engineering and Information Science, University of Science and Technology of China. From 1985 to 1986, she was a Visiting Scientist in the Department of Computer Science, Carnegie Mellon University, Pittsburgh, PA. Her research interests include intelligent information processing, multiresource data fusion, and digital image analysis.

Xian-Sheng Hua received the B.S. and Ph.D. degrees in applied mathematics from Peking University, Beijing, China, in 1996 and 2001, respectively. Since 2001, he has been with Microsoft Research Asia, Beijing, where he is currently a lead researcher with the Media Computing Group. His current research interests include video content analysis, multimedia search, management, authoring, sharing, and advertising. He has authored more than 130 publications in these areas and has more than 30 filed patents or pending applications. He is an adjunct professor at the University of Science and Technology of China, and serves as an associate editor of the IEEE Transactions on Multimedia and an editorial board member of Multimedia Tools and Applications. He won the Best Paper Award and the Best Demonstration Award at ACM Multimedia 2007 and also won the TR35 2008 Young Innovator Award from the MIT Technology Review.