Transductive Multi-Label Learning for Video Concept Detection

Jingdong Wang[†] [*] Yinghai Zhao[‡] Xiuqing Wu[‡] Xian-Sheng Hua[†]
[†]Microsoft Research Asia, Beijing, 100190, P.R.China
[‡]Department of Elec. Eng. & Info. Sci., University of Sci. & Tech. of China, Hefei, 230027, P.R.China
{i-jingdw, xshua}@microsoft.com {yinghai@mail, wuxq@}ustc.edu.cn

ABSTRACT

Transductive video concept detection is an effective way to handle the lack of sufficient labeled videos. However, another issue, the multi-label interdependence, is not essentially addressed in the existing transductive methods. Most solutions only applied the transductive single-label approach to detect each individual concept separately, but ignoring the concept relation, or simply imposed the smoothness assumption over the multiple labels for each video, without indeed exploring the interdependence between the concepts. On the other hand, the semi-supervised extension of supervised multi-label classifiers, such as correlative multi-label support vector machines, is usually intractable and hence impractical due to the quite expensive computational cost. In this paper, we propose an effective transductive multi-label classification approach, which simultaneously models the labeling consistency between the visually similar videos and the multi-label interdependence for each video in an integrated framework. We compare the performance between the proposed approach and several representative transductive single-label 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.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing—*Indexing methods*; I.2.10 [**Artificial Intelligence**]: Vision and Scene Understanding—*Video analysis*

General Terms

Algorithms, theory, experimentation

[*]This work was performed when Yinghai Zhao was visiting Microsoft Research Asia as a research intern.

MIR'08, October 30–31, 2008, Vancouver, British Columbia, Canada.

Keywords

Video concept detection, transductive learning, multi-label interdependence

1. INTRODUCTION

With the rapid growth of digital video content, it becomes more and more of a challenge to manage the videos. A common way is to associate a video with some semantic keywords to describe its semantic content. This poses a challenging topic, video annotation (or video concept detection), which has attracted more and more attention. It builds the relation between the low-level features and the semantic-level concepts to bridge the gap between them. Manual annotation is impractical because it requires intensive labor and much time. Therefore, more attention has been paid on pursuing effective automatic video annotation.

One of the difficulties in semantic video annotation is the insufficiency of the labeled video data. Nowadays, a large-scale video data set can be accessed, but only a few are annotated. The semi-supervised classification technique leveraging the labeled video data as well as the unlabeled video data is adopted to tackle this difficulty. It is expected to obtain a superior performance over the purely-supervised classification method that only utilizes the labeled video data.

In the real world, on the other hand, a video is usually associated with more than one concepts. For example, a video with a "mountain" scene is usually also annotated as an "outdoor" concept. This poses so-called a *multi-label* classification problem, in which a data point may be associated with more than one labels. Most existing semi-supervised classification approaches [19, 22] to video annotation only address the single-label case. Some single-label approaches have been directly applied to multi-label video annotation by transformming it into several independent single-label classification problems.

Recently, some semi-supervised approaches [6, 14, 26] have tried to exploit the multi-label relations. However, they only explore the multi-label inter-similarity, which leads to the label smoothness over multiple labels for each video, i.e., preferring the co-positive or co-negative relations between all pairs of labels. This discourages the cross-positive (i.e., positive-negative or negative-positive) relations and is clearly not accordant to the practice in video annotation. For example, "explosion_fire" and "waterscape_waterfront" seldom occur at the same time. We analyzed the TRECVID 2005 data set and found that the cross-positive relation is dominant on 740 pairs of all the 741 concept pairs among the co-positive and cross-positive relations.

Some supervised methods have been investigated for multi-label classification, e.g., [9, 17]. It is possible to extend them directly to semi-supervised versions through some techniques used in the single-label classification. However, such extensions usually suffer from the heavy computational load, similar to the single label case as pointed out in [2]. Thus it is impractical in the real-world problems, such as video concept detection over the TRECVID data set. This motivates us to investigate an effective and practical approach to semi-supervised multi-label classification.

In this paper, we present a semi-supervised multi-label classification approach based on the discrete hidden Markov random field (MRF) model, which simultaneously models labeling consistency between the visually similar videos and the multi-label interdependence for each video in an integrated framework. It aims to find a labeling such that the multi-label interdependence over unlabeled data points is coherent with that over the labeled ones. We formulate the multi-label interdependence as a pairwise MRF model, which explores all the combinations of relations, including the co-positive, co-negative, and cross-positive relations. Moreover, our approach is very efficient due to the following two aspects. First, it is a trasductive approach and avoids estimating the intermediate inductive decision function. Second, we adopt the efficient graph cuts and tree-reweighted message passing algorithms for labeling inference.

1.1 Related Work

There are many methods for semi-supervised and multi-label classification. This subsection mainly reviews the closely related methods on semi-supervised single-label and multi-label classification, supervised multi-label classification and structured output classification.

A detailed literature survey [28] about semi-supervised single-label classification can be referred to for a comprehensive review. Here, we mainly discuss its application in multimedia. The co-training method has bee applied to video annotation through separating the visual features carefully [18]. Further, the drawbacks of the co-training based video annotation method are analyzed in [23], and an improved co-training style algorithm is proposed. Besides, graph-based transductive classification methods have been widely applied to image and video annotation. The manifold ranking approach is adopted in [8] for image retrieval. A semi-supervised classification method based on kernel density estimation is proposed in [22] for video semantic detection. In addition, several works have been conducted on introducing additional information into the graph-based method, such as the structure cue over the graph [19] and the multi-modality factor [21].

However, less investigation is made to address the multi-label case. There are only a few methods merely explore the multi-label inter-similarity, e.g., [6, 14, 26], which leads to impose the smoothness assumption of the multiple labels for each data point, thus discourage the cross-positive relations and are not accordant to the practice. Moreover, their formulations lead to a semi-definite program or a Sylvester equation, whose computational cost is very expensive.

As aforementioned, it is possible to extend some supervised multi-label classification methods, e.g., [9, 17, 24], to semi-supervised versions. For example, we can extend the state-of-the-art approach, correlative multi-label support vector machines (CML-SVM) [17], through some schemes

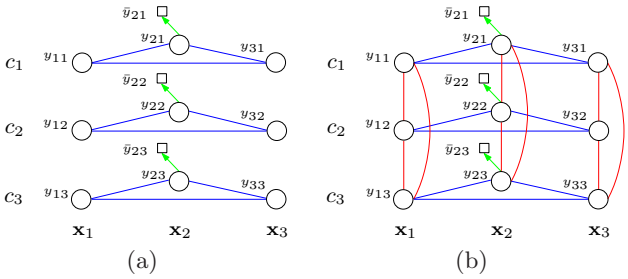


Figure 1: Illustration of the difference between the conventional transductive approach and our approach for multi-label classification. The conventional approach is based on several independent graphs as shown in (a). Our approach considers multi-label interdependence which is depicted in (b).

similar with single-label classifier such as Transductive SVM [10], 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 [10] requires a lot of costly iterations as each iteration need solve a time-consuming CML-SVM over all the data points. From this sense, those semi-supervised extensions are impractical, and hence it is necessary to pursue a practical semi-supervised multi-label approach.

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, 7, 13, 30]. 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 of the multiple nodes, which means to only bias the co-positive or co-negative relations. In [13, 30], 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 without exploring the interdependence of the structured output. The above methods are related to multi-label classification, but 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 can not be 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 propose a discrete hidden Markov random field approach for transductive multi-label classification. Our approach aims to find a labeling, which satisfies two properties similar to 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 is coherent with that over the labeled data, and consists of not only the co-positive and co-negative relations, but also the cross-positive relation.

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 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 has an efficient inference algorithm.
- The multi-label interdependence is modeled by a pairwise Markov random field. All the combinations of pairwise relations over the multiple labels, positive-positive, positive-negative, negative-positive, and negative-negative, are seamlessly modeled in a single integrated formulation.

1.3 Organization

The rest of this paper is organized as follows. Sec. 2 presents the proposed transductive multi-label classification approach and discusses the connections to two kinds of related methods. Then, Sec. 3 reports the comparative experimental results over the TRECVID data set. Finally, Sec. 4 concludes this paper.

2. TRANSDUCTIVE MULTI-LABEL CLASSIFICATION

Suppose there are n data points, $\mathcal{X} = \{\mathbf{x}_i\}_{i\in\mathcal{N}}$, $\mathcal{N} = \{1, \cdots, n\}$, and l points, $\{\mathbf{x}_i\}_{i\in\mathcal{L}}$, are assigned with K-dimensional binary-valued label vectors $\{\bar{\mathbf{y}}_i\}_{i\in\mathcal{L}}, \bar{\mathbf{y}}_i \in \{0,1\}^K$, $\mathcal{L} = \{1, \cdots, l\}$. Here, K is the cardinality of the label set $\mathcal{C} = \{c_1, \cdots, c_K\}$. $\bar{y}_{ic} = 1$ 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 points, $\{\mathbf{x}_i\}_{i\in\mathcal{U}}$, $\mathcal{U} = \{l+1, \cdots, n\}$. Denote $u = n - l$ as the number of unlabeled points, $\mathbf{Y} = [\mathbf{y}_1 \ \cdots \ \mathbf{y}_n]^T$ 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. $\bar{\mathbf{Y}}_l$ 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.

We aim to find a labeling such that it satisfies the three listed properties.

1. It should be the same as the pre-given labeling on the labeled data points.
2. It should be locally smooth, i.e., the labeling should be consistent between the neighboring points.
3. The multi-label interdependence on the unlabeled data points should be similar to that on the labeled data points.

The first two are similar to the previous graph-based transductive single-label method. Differently, we present a discrete formulation which is more natural as the target value in classification is discrete. The last property is novel for semi-supervised multi-label classification.

We present a discrete hidden Markov random field (dHMRF) formulation, which integrates all the three properties into a single framework. In the dHMRF, 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 $\bar{y}_{ic}$. 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 $\mathcal{E}_o = \{(y_{ic}, \bar{y}_{ic})\}_{i\in\mathcal{L}, c\in\mathcal{C}}$. The observable edges correspond to the green edges as shown in Fig. 1(b).

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. We denote this edge set as $\mathcal{E}_s = \{(y_{ic}, y_{jc})\}_{i,j\in\mathcal{L}\cup\mathcal{U}, c\in\mathcal{C}}$. These edges correspond to the blue edges as shown in Fig. 1(b). 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. Such edges are depicted with red edges in Fig. 1(a).

In summary, the constructed graph, $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$, is composed of a node set, $\mathcal{V} = \mathcal{V}_h \cup \mathcal{V}_o$, and an edge set, $\mathcal{E} = \mathcal{E}_o \cup \mathcal{E}_s \cup \mathcal{E}_d$. Here, $\mathcal{E}_o = \cup_{c\in\mathcal{C}}\mathcal{E}_o^c$ is a union of K subsets $\mathcal{E}_o^c$, where each subset is a set of observable edges associated with the concept c. Similarly, $\mathcal{E}_s = \cup_{c\in\mathcal{C}}\mathcal{E}_s^c$, and $\mathcal{E}_d = \cup_{i\in\mathcal{U}}\mathcal{E}_d^i$ with $\mathcal{E}_d^i$ being the set of label-pair-wise edges over the data point $\mathbf{x}_i$. In the following, we mathematically define the potential functions over all the edges on the dHMRF.

2.1 Compatibility with Pre-Labeling

Considering the observable edge $(y_{ic}, \bar{y}_{ic})$ depicted as a green edge in Fig. 1(b), we aim to define a probability between the two variables to describe their compatibility. In our formulation, we fully trust the pre-labeling, and define a discrete probability over the hidden variable y_{ic} as

$$\Pr(y_{ic}) = \delta[y_{ic} = \bar{y}_{ic}], \tag{1}$$

where $\delta[\cdot]$ is an indicator function. This means that y_{ic} must be valued as $\bar{y}_{ic}$ and the probability is zero otherwise. This is essentially equivalent to a *Boltzmann* distribution based on the energy function,

$$e_{i,c}(y_{ic}) = \theta_{i,c}\delta[y_{ic} \neq \bar{y}_{ic}], \tag{2}$$

where $\theta_{i,c}$ corresponds to a penalty that y_{ic} is not equal to $\bar{y}_{ic}$. Here it is valued as ∞ since we expect that the pre-labeling over the labeled data points is unchangeable. $\Pr(y_{ic})$ in Eqn. (1) can be equivalently evaluated as the *Boltzmann* distribution,

$$\Pr(y_{ic}) = \frac{\exp(-e_{i,c}(y_{ic}))}{\exp(-e_{i,c}(0)) + \exp(-e_{i,c}(1))}. \tag{3}$$

Therefore, the joint compatibility is the product of the compatibilities over all the observable edges,

$$\Pr_o(\mathbf{Y}_l) = \prod\nolimits_{i\in\mathcal{L}, c\in\mathcal{C}} \Pr(y_{ic}), \tag{4}$$

which is essentially equivalent to the *Boltzmann* distribution of the loss function,

$$\mathrm{E}_{\mathrm{loss}} = \sum\nolimits_{i\in\mathcal{L}, c\in\mathcal{C}} e_{i,c}(y_{ic}) = \sum\nolimits_{i\in\mathcal{L}, c\in\mathcal{C}} \theta_{i,c}\delta[y_{ic} \neq \bar{y}_{ic}].$$

2.2 Consistency over Local Labeling

As the second property, it is expected that the labeling is consistent between the data points with similar features. It is not straightforward to describe the smoothness of two label vectors, i.e., the labeling configurations over two data points. Instead, we factor the smoothness between two label vectors into an aggregation of the smoothness between all the pairs of corresponding entries. Specifically, we formulate it as

$$\psi(\mathbf{y}_i, \mathbf{y}_j) = \prod\nolimits_{c\in\mathcal{C}} \psi(y_{ic}, y_{jc}). \quad (5)$$

Here the potential function over y_{ic} and y_{jc} is defined as

$$\psi(y_{ic}, y_{jc}) = \sum\nolimits_{p,q\in\{0,1\}} P^{pq}_{ij,c}\delta[y_{ic} = p \wedge y_{jc} = q], \quad (6)$$

where $\sum_{p,q\in\{0,1\}} P^{pq}_{ij,c} = 1$, and $P^{pq}_{ij,c}$ is the probability when $y_{ic} = p$ and $y_{jc} = q$. We define it as a *Boltzmann* distribution of an Ising energy over an interacting pair,

$$e_{ij,c}(y_{ic}, y_{jc}) = \theta^0_{ij,c}\delta[y_{ic} \neq y_{jc}] + \theta^1_{ij,c}\delta[y_{ic} = y_{jc}], \quad (7)$$

where $\theta^0_{ij,c}$ is a penalty that only one of $\mathbf{x}_i$ and $\mathbf{x}_j$ is assigned the label c, and $\theta^1_{ij,c}$ is a penalty that both $\mathbf{x}_i$ and $\mathbf{x}_j$ are assigned or not assigned the label c. Similar to [29], we define the two penalties 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 $\theta^1_{ij,c} = 0$ and $\theta^0_{ij,c} = w_{ij}$, $w_{ij} = \exp(-\gamma\|\mathbf{x}_i - \mathbf{x}_j\|^2)$ with $\gamma > 0$ being a kernel parameter.

Combining all the potentials over the sample-pair-wise edges, the aggregated probability is formulated as

$$\begin{aligned}\Pr_s(\mathbf{Y}) &= \frac{1}{Z_s}\prod\nolimits_{(i,j)\in\mathcal{E}_s} \psi(\mathbf{y}_i, \mathbf{y}_j) \\ &= \frac{1}{Z_s}\prod\nolimits_{(i,j)\in\mathcal{E}^c_s, c\in\mathcal{C}} \psi(y_{ic}, y_{jc}), \end{aligned} \quad (8)$$

which is essentially as well the *Boltzmann* distribution of the regularization function,

$$\mathrm{E}_{\mathrm{reg}} = \sum\nolimits_{c\in\mathcal{C}} \mathrm{E}_{\mathrm{reg},c} = \sum\nolimits_{(i,j)\in\mathcal{E}^c_s, c\in\mathcal{C}} \theta^0_{ij,c}\delta[y_{ic} \neq y_{jc}]. \quad (9)$$

2.3 Interdependence among Multiple Labels

Some existing semi-supervised methods, such as in [19, 21], 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,

$$\Pr(\mathbf{Y}) = \prod\nolimits_{c\in\mathcal{C}} \Pr(\mathbf{y}^c). \quad (10)$$

This factorization is equivalent to defining a probability over the K isolated subgraphs, $\{\mathcal{G}_1, \cdots, \mathcal{G}_K\}$, 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 this paper, we consider the pairwise interdependence between the multiple labels, which is modeled as a pairwise Markov random field formulation. Specifically, we define the interdependence probability as a product of the potential functions over all the pairs of labels for each data point,

$$\Pr(\mathbf{y}_i) = \frac{1}{Z'}\prod\nolimits_{\alpha,\beta\in\mathcal{C}} \phi(y_{i\alpha}, y_{i\beta}), \quad (11)$$

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

$$\phi(y_{i\alpha}, y_{i\beta}) = \sum\nolimits_{p,q\in\{0,1\}} P^{pq}_{\alpha\beta}\delta[y_{i\alpha} = p \wedge y_{i\beta} = q]. \quad (12)$$

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

$$\prod\nolimits_{i\in\mathcal{L}} \Pr(\mathbf{y}_i) = \prod\nolimits_{i\in\mathcal{L},\alpha,\beta\in\mathcal{C}} \frac{1}{Z'}\phi(y_{i\alpha}, y_{i\beta}). \quad (13)$$

The estimation is NP-hard since our MRF is a loopy graph. In [12], 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,

$$P^{pq}_{\alpha\beta} = \frac{1}{|\mathcal{L}|}\sum\nolimits_{i\in\mathcal{L}} \delta[y_{i\alpha} = p \wedge y_{i\beta} = q], \quad (14)$$

where $p, q \in \{0, 1\}$, and $\delta[\cdot]$ 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, $P^{pq}_{\alpha\beta} = P^{pq}_{\alpha\beta} + \epsilon$, where ϵ is a small constant that is valued by $\frac{100}{n}$ 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 jointed together as

$$\Pr_d(\mathbf{Y}_u) = \prod\nolimits_{i\in\mathcal{U}} \Pr(\mathbf{y}_i) = \frac{1}{Z_d}\prod\nolimits_{i\in\mathcal{U},\alpha,\beta\in\mathcal{C}} \phi(y_{i\alpha}, y_{i\beta}),$$

where Z_d is a normalization constant.

It is worth pointing out that the multi-label interdependence that we take into accounts consists of all possible relations between the concept pairs. In [26], 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 cross-positive 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 cross-positive relations, including the negative-positive and positive-negative relations, are reasonably captured.

2.4 Overall Objective

The above three probabilities are aggregated into a single probability,

$$\begin{aligned}\Pr(\mathbf{Y}) &= \frac{1}{Z}\Pr\nolimits_o(\mathbf{Y}_l)\Pr\nolimits^{\lambda}_s(\mathbf{Y})\Pr\nolimits^{1-\lambda}_d(\mathbf{Y}_u) \\ &= \frac{1}{Z}\prod\nolimits_{i\in\mathcal{L},c\in\mathcal{C}} \Pr(y_{ic})\prod\nolimits_{c\in\mathcal{C},(i,j)\in\mathcal{E}^c_s} \psi^{\lambda}(y_{ic}, y_{jc}) \\ &\quad \prod\nolimits_{i\in\mathcal{U},\alpha,\beta\in\mathcal{C}} \phi^{1-\lambda}(y_{i\alpha}, y_{i\beta}), \end{aligned} \quad (15)$$

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 that is defined over the green edge. The second term corresponds to the label consistency between the data points with similar features that is defined over the blue edge. The last term corresponds to the multi-label interdependence that is defined over the red edge. The solution to the proposed transductive multi-label classification is found as the joint maximum,

$$\mathbf{Y}^* = \arg\max_{\mathbf{Y}} \Pr(\mathbf{Y}). \tag{16}$$

For efficiently labels inferring, we combine tree-reweighted message passing [11] and graph cuts [4] to optimize the discrete hidden Markov random field. Tree-reweighted message passing can obtain a near-optimal solution but its speed depends on the initialization. Graph cuts is guaranteed to obtain the global optimal solution but the pairwise energy function must be submodular. Therefore, we discard the non-submodular energies and perform the graph cuts algorithm to provide a good initialization, which is used as the warm start, to speed up the inference.

2.5 Concept Scoring

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

$$\begin{aligned} & \Pr(y_{ic} = 1 | \mathbf{Y}_{\setminus y_{ic}} = \mathbf{Y}^*_{\setminus y_{ic}}) \\ = \; & \Pr(y_{ic} = 1 | y_{jc} = y^*_{jc}, (i,j) \in \mathcal{E}^c_s, y_{i\beta} = y^*_{i\beta}, \beta \in \mathcal{C} \setminus \{c\}) \\ = \; & \frac{1}{Z} \prod\nolimits_{(i,j)\in\mathcal{E}^c_s} \psi^{\lambda}(y_{ic} = 1, y_{jc} = y^*_{jc}) \\ & \prod\nolimits_{\beta\in\mathcal{C}\setminus\{c\}} \psi^{1-\lambda}(y_{ic} = 1, y_{i\beta} = y^*_{i\beta}), \end{aligned} \tag{17}$$

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

2.6 Discussion

We analyze the connections between our approach and several related methods: the Gaussian random field (GRF) approach [29] and the Sylvester equation based approach [26].

For connection with GRF, if we only take the regularization $\mathrm{E}_{\mathrm{reg},c}$ over single label c in Eqn. (9) into account, considering $y_i \in \{0, 1\}$ and $\theta^1_{ij} = 0$ in Eqn. (7), the regularization is transformed to the following discrete function,

$$\mathrm{E}_{\mathrm{reg}} = \sum\nolimits_{ij} \theta^0_{ij} (y_i - y_j)^2. \tag{18}$$

Suppose $\theta^0_{ij} = w_{ij}$ and discard the integer constraint $y_i \in \{0, 1\}$, GRF is a degradation of our approach.

For connection with Sylvester equation, if we transform the multi-label interdependence probability in Eqn. (12) to the equivalent energy formulation and combine all points' energy functions in matrix formulation, a Sylvester equation can be obtained by relaxing discrete labels into real numbers.

$$\lambda \mathbf{A}\mathbf{Y}_u + (1-\lambda)\mathbf{Y}_u\mathbf{B} = -\lambda\mathbf{C}\bar{\mathbf{Y}}_l, \tag{19}$$

where $\mathbf{Y}_u$ and $\bar{\mathbf{Y}}_l$ 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 [26] is also a degraded version of our approach.

3. EXPERIMENTS

3.1 Setup

We evaluate the proposed transductive multi-label classification arpproach 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 [20]. Following the data set separation scheme presented in [25], 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. For each sub-shot, we extract a low-level feature from the corresponding key-frame. In our experiments, we use a 225-dimensional block-wise color moment, which is extracted over 5×5 blocks with each block described by a 9-dimensional feature, and normalize each dimension with a standard normal distribution.

According to the LSCOM-Lite annotations [15], 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. As pointed in [17], 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.

3.2 Comparative Experiments

In our experiments, we perform two types of comparative experiments with the popular graph-based transductive single-label classification methods and the state of the art of supervised multi-label classification.

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-pairwise edges are constructed by connecting a video and its 30 nearest neighbors. To speed up the labeling inference, we group 36 of all the concepts into 3 subsets, called label chunklets, by spectral clustering [16] according to their interdependence. Meanwhile, the remaining three concepts, "face", "outdoor" and "person", which have strong interdependence with all the other concepts, are added to the three concept chunklets forming three augmented chunklets. Intuitively, only the interdependence between the labels from the same chunklets is significant, and the dependence between the labels from different chunklets is very minor (except the three common concepts) and hence neglected. The concepts in each augmented chunklet are connected with

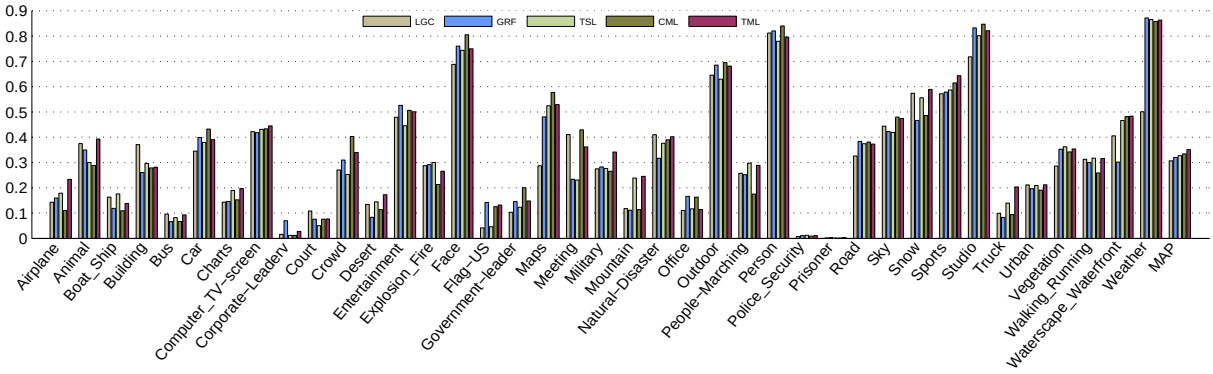


Figure 2: Comparison with transductive single-label classification and supervised multi-Label classification.

each other to form the label-pair-wise edges. Therefore, we obtain three isolated graphs. 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. The kernel parameter γ and the trade-off variable λ are obtained by validation.

3.2.1 Vs. Transductive Single-Label Classification

In this experiment, we compare the proposed approach with two popular graph-based transductive single-label classification approaches: the Gaussian random field (GRF) [29] and the local and global consistency (LGC) method [27].

The graph for GRF and LGC 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. The kernel parameter is selected through a similar validation process.

In addition, we conduct an experiment on the 39 concept 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 scoring schemes. In the experiment, we denote this method by TSL.

The comparative results are shown in Fig. 2, from which it can be observed that multi-label interdependence is capable to improve 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% and 9.7% relative improvements, compared with LGC and GRF. 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 and 26 of all the 39 concepts than LGC and GRF. 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.
- 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 is 1% deterioration on "face", 0.6% deterioration on "outdoor", and 3% 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.

3.2.2 Vs. Supervised Multi-Label Classification

In this experiment, we compare the proposed approach with the state of the art of supervised multi-label classification proposed in [17], correlative multi-label classification (CML). For CML, we adopt the Gaussian kernel and the kernel parameter and the trade-off variable are obtained through the validation process, which is the same with [17].

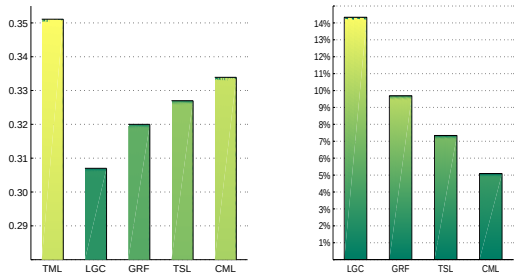
The comparative results are shown in Fig. 2. We can see that our approach outperforms CML. This superiority mainly comes from the transductive scheme, which directly estimates the labeling without estimating an intermediate classifier. The following observation can be obtained.

- TML obtains a 5.1% relative improvement over CML.
- TML outperforms CML on 26 of 39 concepts. For some concepts, TML obtains significant improvements. For example, 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" and "studio", than CML. This deterioration is due to the local labeling consistency constraint, which smoothes some too high AP scores.

The MAPs of the above existing approaches and the improvements of our approach over the other methods are summarized in Fig. 3. It can be seen that our approach get the best overall performance and the improvements are notable.

4. CONCLUSIONS

In this paper, we propose a novel transductive multi-label learning approach for video concept detection. Different from the existing approaches, the proposed approach simultaneously models the labeling consistency between visually-similar videos and the multi-label interdependence for each video in an integrated framework. We formulate the multi-label interdependence as a pairwise Markov random field model and learn it from the labeled data points. The superiority of the performance over several existing approaches on the TRECVID data set shows its effectiveness.



(a) MAP (b) Improvement rate

Figure 3: (a) shows the MAPs of several existing methods, and (b) indicates the improvement rates of our approach over other methods.

5. REFERENCES

[1] Y. Altun, D. A. McAllester, and M. Belkin. Maximum Margin Semi-Supervised Learning for Structured Variables. In *NIPS*, 2005.

[2] M. Belkin, P. Niyogi, and V. Sindhwani. Manifold Regularization: A Geometric Framework for Learning from Labeled and Unlabeled Examples. *Journal of Machine Learning Research*, 7:2399–2434, 2006.

[3] K. P. Bennett and A. Demiriz. Semi-Supervised Support Vector Machines. In *NIPS*, pages 368–374, 1998.

[4] Y. Boykov and V. Kolmogorov. An Experimental Comparison of Min-Cut/Max-Flow Algorithms for Energy Minimization in Vision. *IEEE Trans. Pattern Anal. Mach. Intell.*, 26(9):1124–1137, 2004.

[5] U. Brefeld and T. Scheffer. Semi-Supervised Learning for Structured Output Variables. In *ICML*, pages 145–152, 2006.

[6] G. Chen, Y. Song, F. Wang, and C. Zhang. Semi-Supervised Multi-Label Learning by Solving a Sylvester Equation. In *SDM*, 2008.

[7] K. Duh and K. Kirchhoff. Structured Multi-Label Transductive Learning. In *NIPS Workshop on Advances in Structured Learning for Text/Speech Processing*, 2005.

[8] J. He, M. Li, H. Zhang, H. Tong, and C. Zhang. Manifold-Ranking based Image Retrieval. In *ACM Multimedia*, pages 9–16, 2004.

[9] W. Jiang, S.-F. Chang, and A. C. Loui. Active Context-Based Concept Fusion with Partial User Labels. In *ICIP*, pages 2917–2920, 2006.

[10] T. Joachims. Transductive Inference for Text Classification using Support Vector Machines. In *ICML*, pages 200–209, 1999.

[11] V. Kolmogorov. Convergent Tree-Reweighted Message Passing for Energy Minimization. *IEEE Trans. Pattern Anal. Mach. Intell.*, 28(10):1568–1583, 2006.

[12] H. H. Ku and S. Kullback. Approximating Discrete Probability Distributions. *IEEE Transactions on Information Theory*, IT-15(4):444–447, 1969.

[13] C.-H. Lee, S. Wang, F. Jiao, D. Schuurmans, and R. Greiner. Learning to Model Spatial Dependency: Semi-Supervised Discriminative Random Fields. In *NIPS*, pages 793–800, 2006.

[14] Y. Liu, R. Jin, and L. Yang. Semi-supervised Multi-label Learning by Constrained Non-Negative Matrix Factorization. In *AAAI*, 2006.

[15] M. R. Naphade, L. Kennedy, J. R. Kender, S.-F. Chang, J. R. Smith, P. Over, and A. Hauptmann. A Light Scale Concept Ontology for Multimedia Understanding for TRECVID 2005. In *IBM Research Report RC23612 (W0505-104)*, 2005.

[16] A. Y. Ng, M. I. Jordan, and Y. Weiss. On Spectral Clustering: Analysis and an Algorithm. In *NIPS*, pages 849–856, 2001.

[17] G.-J. Qi, X.-S. Hua, Y. Rui, J. Tang, T. Mei, and H.-J. Zhang. Correlative Multi-Label Video Annotation. In *ACM Multimedia*, pages 17–26, 2007.

[18] Y. Song, X.-S. Hua, L.-R. Dai, and M. Wang. Semi-Automatic Video Annotation based on Active Learning with Multiple Complementary Predictors. In *Multimedia Information Retrieval*, pages 97–104, 2005.

[19] J. Tang, X.-S. Hua, G.-J. Qi, M. Wang, T. Mei, and X. Wu. Structure-Sensitive Manifold Ranking for Video Concept Detection. In *ACM Multimedia*, pages 852–861, 2007.

[20] TRECVID2005. http://www-nlpir.nist.gov/projects/trecvid/.

[21] M. Wang, X.-S. Hua, X. Yuan, Y. Song, and L.-R. Dai. Optimizing Multi-Graph Learning: towards a Unified Video Annotation Scheme. In *ACM Multimedia*, pages 862–871, 2007.

[22] M. Wang, Y. Song, X. Yuan, H. Zhang, X.-S. Hua, and S. Li. Automatic Video Annotation by Semi-Supervised Learning with Kernel Density Estimation. In *ACM Multimedia*, pages 967–976, 2006.

[23] R. Yan and M. R. Naphade. Semi-Supervised Cross Feature Learning for Semantic Concept Detection in Videos. In *CVPR (1)*, pages 657–663, 2005.

[24] R. Yan, M. yu Chen, and A. G. Hauptmann. Mining Relationship Between Video Concepts using Probabilistic Graphical Models. In *ICME*, pages 301–304, 2006.

[25] A. Yanagawa, S.-F. Chang, L. Kennedy, and W. Hsu. Columbia University's Baseline Detectors for 374 LSCOM Semantic Visual Concepts. Technical report, Columbia University, March 2007.

[26] Z.-J. Zha, T. Mei, J. Wang, Z. Wang, and X.-S. Hua. Graph-Based Semi-Supervised Learning with Multi-Label. In *ICME*, pages 1321–1324, 2008.

[27] D. Zhou, O. Bousquet, T. N. Lal, J. Weston, and B. Schölkopf. Learning with Local and Global Consistency. In *NIPS*, 2003.

[28] X. Zhu. Semi-supervied Learning Literature Survey. *Computer Sciences Technical Report, 1530, University of Wisconsin-Madison*, 2007.

[29] X. Zhu, Z. Ghahramani, and J. Lafferty. Semi-Supervised Learning Using Gaussian Fields and Harmonic Functions. In *ICML*, pages 912–919, 2003.

[30] A. Zien, U. Brefeld, and T. Scheffer. Transductive Support Vector Machines for Structured Variables. In *ICML*, pages 1183–1190, 2007.