Correlative Multi-Label Video Annotation

Guo-Jun Qi[*]
Dept. of Automation
Univ. of Sci. & Tech. of China
Hefei, Anhui, 230027 China
qgj@mail.ustc.edu.cn

Xian-Sheng Hua
Microsoft Research Asia
49 Zhichun Road
Beijing, 100080 China
xshua@microsoft.com

Yong Rui
Microsoft China R&D Group
49 Zhichun Road
Beijing, 100080 China
yongrui@microsoft.com

Jinhui Tang[*]
Dept. of Elec. Eng. & Info. Sci.
Univ. of Sci. & Tech. of China
Hefei, Anhui, 230027 China
jhtang@mail.ustc.edu.cn

Tao Mei
Microsoft Research Asia
49 Zhichun Road
Beijing, 100080 China
tmei@microsoft.com

Hong-Jiang Zhang
Microsoft Adv. Tech. Center
49 Zhichun Road
Beijing, 100080 China
hjzhang@microsoft.com

ABSTRACT

Automatically annotating concepts for video is a key to semantic-level video browsing, search and navigation. The research on this topic evolved through two paradigms. The first paradigm used binary classification to detect each individual concept in a concept set. It achieved only limited success, as it did not model the inherent correlation between concepts, e.g., urban and building. The second paradigm added a second step on top of the individual-concept detectors to fuse multiple concepts. However, its performance varies because the errors incurred in the first detection step can propagate to the second fusion step and therefore degrade the overall performance. To address the above issues, we propose a third paradigm which simultaneously classifies concepts and models correlations between them in a single step by using a novel *Correlative Multi-Label* (CML) framework. We compare the performance between our proposed approach and the state-of-the-art approaches in the first and second paradigms on the widely used TRECVID data set. We report superior performance from 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 G.-J. Qi and J. Tang were visiting Microsoft Research Asia as research interns.

Keywords

Video Annotation, Multi-Labeling, Concept Correlation

1. INTRODUCTION

Automatically annotating video at the semantic concept level has emerged as an important topic in the multimedia research community [11][16]. The concepts of interest include a wide range of categories such as scenes (e.g., urban, sky, mountain, etc.), objects (e.g., airplane, car, face, etc.), events (e.g., explosion-fire, people-marching, etc.) and certain named entities (e.g. person, place, etc.) [16][12]. Before we discuss the details of this topic, we would like to first define a few terminologies. The annotation problem of interest to this paper, as well as to other research efforts [16][12], is a *multi-labeling* process where a video clip can be annotated with multiple labels. For example, a video clip can be classified as "*urban*", "*building*" and "*road*" simultaneously. In contrast, *multi-class* annotation process labels only one concept to each video clip. Most of the real-world problems, e.g., the ones being addressed in TRECVID [17], are multi-label annotation, not multi-class annotation. In addition, multi-label is more complex and challenging than multi-class, as it involves non-exclusive detection and classification. This paper focuses on multi-label annotation.

Research on multi-label video annotation evolved through two paradigms: individual concept detection and annotation, and *Context Based Conceptual Fusion* (CBCF) [8] annotation. In this paper, we propose the third paradigm: integrated multi-label annotation. We next review these three paradigms.

1.1 First Paradigm: Individual Concept Annotation

In this paradigm, multiple video concepts are detected *individually* and *independently* without considering correlations between them. That is, the multi-label video annotation is translated into a set of binary detectors with presence/absence of the label for each concept. A typical approach is to independently train a concept model using *Support Vector Machine*(SVM) [4] or *Maximum Entropy Model* (MEM) [13] etc. The leftmost flowchart of Figure 1 illustrates the first paradigm – a set of individual SVMs for video concept detection and annotations. A mathematical alternative is to stack this set of detectors into a single discrimi-
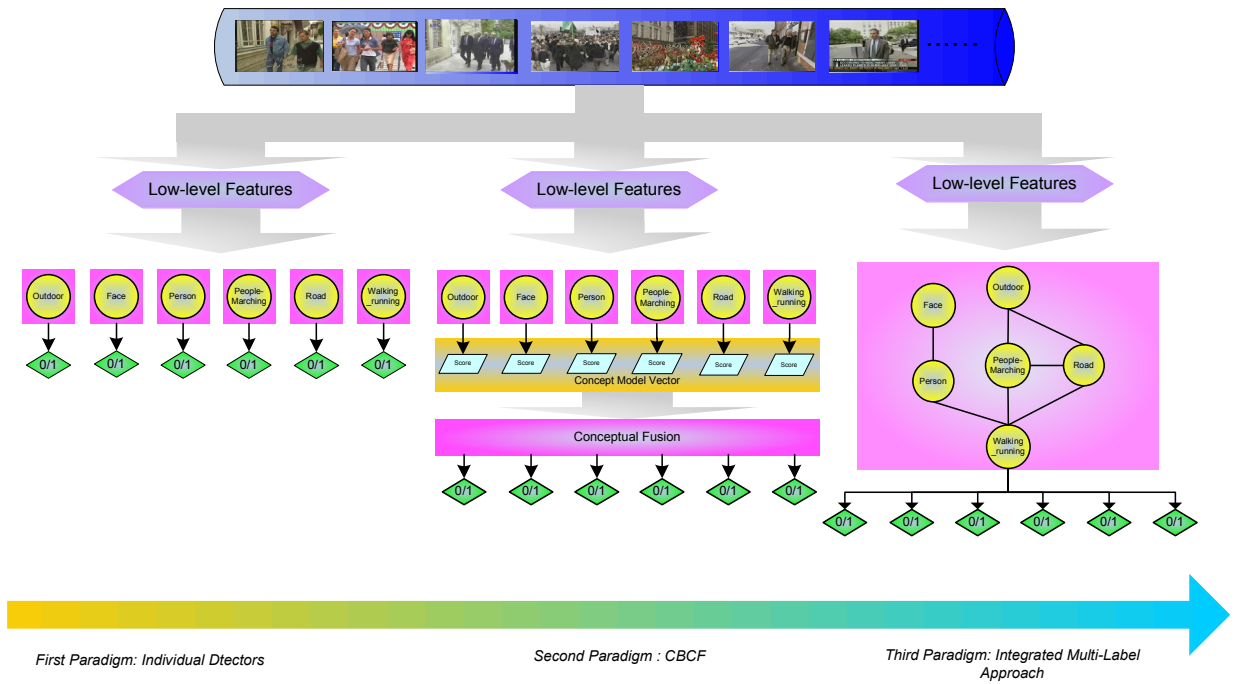
Figure 1: The multi-label video annotation methods in three paradigms. From leftmost to the rightmost, they are the individual SVM, CBCF and our proposed CML.

native classifier [14]. However, both the individual detectors and the stacked classifier at their core are independent binary classification formulations.

The first-paradigm approaches only achieved limited success. In real world, video concepts do not exist in isolation. Instead, they appear correlatively and naturally interact with each other at the semantic level. For example, the presence of "crowd" often occurs together with the presence of "people" while "boat ship" and "truck" commonly do not co-occur. Furthermore, while simple concepts can be modeled directly from low level features, it is quite difficult to individually learn the models of complex concepts, e.g., "people marching", from the low-level features. Instead, the complex concepts can be better inferred based on the correlations with the other concepts. For instance, the presence of "people marching" can be boosted if both "crowd" and "walking running" occurs in a video clip.

1.2 Second Paradigm: Context Based Conceptual Fusion Annotation

One of the most well-known approaches in this paradigm is to refine the detection results of the individual detectors with a *Context Based Concept Fusion* (CBCF) strategy. For instance, Naphade et al. [10] proposed a probabilistic Bayesian Multinet approach to explicitly model the relationship between the multiple concepts through a factor graph which is built upon the underlying video ontology semantics. Wu et al. [20] used an ontology-based multi-classification learning for video concept detection. Each concept is first independently modeled by a classifier, and then a predefined ontology hierarchy is investigated to improve the detection accuracy of the individual classifiers. Smith et al. [15] presented a two-step Discriminative Model Fusion (DMF) ap-

proach to mine the unknown or indirect relationship to specific concepts by constructing model vectors based on detection scores of individual classifiers. A SVM is then trained to refine the detection results of the individual classifiers. The center flowchart of Figure 1 shows such a second-paradigm approach. Alternative fusion strategy can also be used, e.g. Hauptmann et al. [6] proposed to use Logistic Regression (LR) to fuse the individual detections. Jiang et al. [8] used a CBCF-based active learning method. Users were involved in their approach to annotate a few concepts for extra video clips, and these manual annotations were then utilized to help infer and improve detections of other concepts.

Although it is intuitively correct that contextual relationship can help improve detection accuracy of individual detectors, experiments of the above CBCF approaches have shown that such improvement is not always stable, and the overall performance can even be worse than individual detectors alone. For example, in [6] at least 3 out of 8 concepts do not gain better performance by using the conceptual fusion with a LR classifier atop the uni-concept detectors. The unstable performance gain is due to the following reasons:

1. CBCF methods are built on top of the independent binary detectors with a second step to fuse them. However, the output of the individual independent detectors can be unreliable and therefore their detection errors can propagate to the second fusion step. As a result, the final annotations can be corrupted by these incorrect predictions. From a philosophical point of view, the CBCF approaches do not follow the *principle of Least-Commitment* espoused by D. Marr [9], because they are prematurely committed to irreversible individual predictions in the first step which can or cannot be corrected in the second fusion step.

2. A secondary reason comes from the insufficient data for the conceptual fusion. In CBCF methods, the samples needs to be split into two parts for each step and the samples for the conceptual fusion step is usually insufficient compared to the samples used in the first training step. Unfortunately, the correlations between the concepts are usually complex, and insufficient data can lead to "over fitting" in the fusion step, thus the obtained prediction lacks the generalization ability.

1.3 Third Paradigm: Integrated Multi-label Annotation

To address the difficulties faced in the first and second paradigms, in this paper, we will propose a third paradigm. The key of this paradigm is to simultaneously model both the individual concepts and their interactions in a single formulation. The rightmost flowchart of Figure 1 illustrates our proposed *Correlative Multi-Label* (CML) method. This approach has the following advantages compared with the second paradigm, e.g., CBCF methods:

1. The approach follows the *Principle of Least-Commitment* [9]. Because the learning and optimization is done in a single step for all the concepts simultaneously, it does not have the error propagation problem as in CBCF.

2. The entire samples are efficiently used simultaneously in modeling the individual concepts as well as their correlations. The risk of overfitting due to the insufficient samples used for modeling the conceptual correlations is therefore significantly reduced.

To summarize, the first paradigm does not address concept correlation. The second paradigm attempts to address it by introducing a separate second correlation step. The third paradigm, on the other hand, addresses the correlation issue at the root in a single step. The rest of the paper is organized as follows. In Section 2, we give a detailed description of the proposed *Correlative Multi-Label* (CML) approach, including the classification model, and the learning strategy. In Section 3, we will explore the connection between the proposed approach and *Gibbs Random Fields* (GRFs) [19], based on which we can show an intuitive interpretation on how the proposed approach captures the individual concepts as well as the conceptual correlations. Section 4 details the implementation issues, including concept label vector prediction and concept scoring. Finally, in Section 5, we will report experiments on the benchmark TRECVID data and show that the proposed approach has superior performance over state-of-the-art algorithms in both first and second paradigms.

2. OUR APPROACH-CML

In this section, we will introduce our proposed correlative multi-labeling (CML) model for video semantic annotation. In Section 2.1, we will present the mathematical formulation of the multi-labeling classification function, and show that this function captures the correlations between the individual concepts and low-level features, as well as the correlations between the different concepts. Then in Section 2.2, we will describe the learning procedure of the proposed CML model.

2.1 A Multi-Label Classification Model

Let $\boldsymbol{x} = (x_1, x_2, \cdots, x_D)^T \in \mathcal{X}$ denote the input pattern representing feature vectors extracted from video clips; Let

$\boldsymbol{y} \in \mathcal{Y} = \{+1, -1\}^K$ denote the K dimensional concept label vector of an example, where each entry $y_i \in \{+1, -1\}$ of $\boldsymbol{y}$ indicates the membership of this example in the *ith* concept. $\mathcal{X}$ and $\mathcal{Y}$ represent the input feature space and label space of the data set, respectively. The proposed algorithm aims at learning a linear discriminative function

$$F(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{w}) = \langle \boldsymbol{w}, \theta(\boldsymbol{x}, \boldsymbol{y}) \rangle \qquad (1)$$

where $\theta(\boldsymbol{x}, \boldsymbol{y})$ is a vector function mapping from $\mathcal{X} \times \mathcal{Y}$ to a new feature vector which encodes the models of individual concepts as well as their correlations together (to be detailed later); $\boldsymbol{w}$ is the linear combination weight vector. With such a discriminative function, for an input pattern $\boldsymbol{x}$, the label vector $\boldsymbol{y}^*$ can be predicted by maximizing over the argument $\boldsymbol{y}$ as

$$\boldsymbol{y}^* = \max_{y \in \mathcal{Y}} F(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{w}) \qquad (2)$$

As to be presented in the next section, such a discriminative function can be intuitively interpreted in the Gibbs random fields (GRFs) [19] framework when considering the defined feature vector $\theta(\boldsymbol{x}, \boldsymbol{y})$. The constructed feature $\theta(\boldsymbol{x}, \boldsymbol{y})$ is a high-dimensional feature vector, whose elements can be partitioned into two types as follows. And as to be shown later these two types of elements actually account for modeling of individual concepts and their interactions, respectively.

Type I The elements for *individual* concept modeling:

$$\theta_{d,p}^l(\boldsymbol{x}, \boldsymbol{y}) = x_d \cdot \delta \llbracket y_p = l \rrbracket, \\ l \in \{+1, -1\}, 1 \le d \le D, 1 \le p \le K \qquad (3)$$

where $\delta \llbracket y_p = l \rrbracket$ is an indicator function that takes on value 1 if the predict is true and 0 otherwise; D and K are the dimensions of low level feature vector space $\mathcal{X}$ and the number of the concepts respectively. These entries of $\theta(\boldsymbol{x}, \boldsymbol{y})$ serve to model the connection between the low level feature $\boldsymbol{x}$ and the labels $y_k (1 \le k \le K)$ of the concepts. They have the similar functionality as in the traditional SVM which models the relations between the low-level features and high-level concepts.

However, as we have discussed, it is not enough for a multi-labeling algorithm to only account for modeling the connections between the labels and low-level features without considering the semantic correlations of different concepts. Therefore, another element type of $\theta(\boldsymbol{x}, \boldsymbol{y})$ is required to investigate the correlations between the different concepts.

Type II The elements for concept correlations:

$$\theta_{p,q}^{m,n}(\boldsymbol{x}, \boldsymbol{y}) = \delta \llbracket y_p = m \rrbracket \cdot \delta \llbracket y_q = n \rrbracket \\ m, n \in \{+1, -1\}, 1 \le p < q \le K \qquad (4)$$

where the superscripts m and n are the binary labels (positive and negative label), and subscripts p and q are the concept indices. These elements serve to capture all the possible pairs of concepts and labels. Note that, both positive and negative relations are captured with these elements. For example, the concept "building" and "urban" is a positive concept pair that often co-occurs while "explosion fire" and "waterscape waterfront" is negative concept pair that usually does not occur at the same time.

Note that we can model high-order correlations among these concepts as well, but it will require more training samples. As to be shown in Section 5, such an order-2 model successfully trades off between the model complexity and concept correlation complexity, and achieves significant improvement in the concept detection performance.

We concatenate the above two types of elements to form the feature vector $\theta(\boldsymbol{x}, \boldsymbol{y})$. It is not difficult to see that the dimension of vector $\theta(\boldsymbol{x}, \boldsymbol{y})$ is $2KD + 4C_K^2 = 2K(D+K-1)$. When K and D are large, the dimension of $\theta(\boldsymbol{x}, \boldsymbol{y})$ will be extraordinary high. For example, if $K = 39$ and $D = 200$, $\theta(\boldsymbol{x}, \boldsymbol{y})$ will have $18,564$ dimensions. However, this vector is *sparse* thanks to the indicator function $\delta \llbracket \cdot \rrbracket$ in Eqns. (3) and (4). This is a key step in the mathematical formulation. As a result, the kernel function (i.e. the dot product) between the two vectors, $\theta(\boldsymbol{x}, \boldsymbol{y})$ and $\theta(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}})$, can be represented in a very compact form as

$$\begin{aligned}\langle \theta(\boldsymbol{x}, \boldsymbol{y}), \theta(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}}) \rangle = \langle \boldsymbol{x}, \tilde{\boldsymbol{x}} \rangle \sum_{1 \le k \le K} \delta \llbracket y_k = \tilde{y}_k \rrbracket \\ + \sum_{1 \le p < q \le K} \delta \llbracket y_p = \tilde{y}_p \rrbracket \delta \llbracket y_q = \tilde{y}_q \rrbracket\end{aligned} \quad (5)$$

where $\langle \boldsymbol{x}, \tilde{\boldsymbol{x}} \rangle$ is the dot product over the low-level feature vector x and $\tilde{\boldsymbol{x}}$. Of course, a Mercer kernel function $K(\boldsymbol{x}, \tilde{\boldsymbol{x}})$ (such as Gaussian Kernel, Polynomial Kernel) can be substituted for $\langle \boldsymbol{x}, \tilde{\boldsymbol{x}} \rangle$ as in the conventional SVMs, and *nonlinear* discriminative functions can then be introduced with the use of these kernels. In the next subsection, we will present the learning procedure of this model. As to be described, the above compact kernel representation will be used explicitly in the learning procedure instead of the original feature vector $\theta(\boldsymbol{x}, \boldsymbol{y})$.

2.2 Learning the Classifier

Using the feature vector we constructed above and its kernel representation in (5), the learning procedure trains a classification model as delineated in (1). The procedure follows a similar derivation as in the conventional SVM (details about SVM can be found in [4]) and in particular one of its variants for the structural output spaces [18]. Given an example $\boldsymbol{x}_i$ and its label vector $\boldsymbol{y}_i$ from the training set $\{\boldsymbol{x}_i, \boldsymbol{y}_i\}_{i=1}^n$, according to Eqn. (1) and (2), a misclassification occurs when we have

$$\begin{aligned}\Delta F_i(\boldsymbol{y}) &\triangleq F(\boldsymbol{x}_i, \boldsymbol{y}_i) - F(\boldsymbol{x}_i, \boldsymbol{y}) \\ &= \langle \boldsymbol{w}, \Delta \theta_i(\boldsymbol{y}) \rangle \le 0, \forall \boldsymbol{y} \ne \boldsymbol{y}_i, \boldsymbol{y} \in \mathcal{Y}\end{aligned} \quad (6)$$

where $\Delta \theta_i(\boldsymbol{y}) = \theta(\boldsymbol{x}_i, \boldsymbol{y}_i) - \theta(\boldsymbol{x}_i, \boldsymbol{y})$. Therefore, the empirical prediction risk on training set wrt the parameter $\boldsymbol{w}$ can be expressed as

$$\hat{R}(\{\boldsymbol{x}_i, \boldsymbol{y}_i\}_{i=1}^n; \boldsymbol{w}) = \frac{1}{n} \sum_{i=1}^n \sum_{\boldsymbol{y} \ne \boldsymbol{y}_i, \boldsymbol{y} \in \mathcal{Y}} \ell(\boldsymbol{x}_i, \boldsymbol{y}; \boldsymbol{w}) \quad (7)$$

where $\ell(\boldsymbol{x}_i, \boldsymbol{y}; \boldsymbol{w})$ is a loss function counting the errors as

$$\ell(\boldsymbol{x}_i, \boldsymbol{y}; \boldsymbol{w}) = \left\{ \begin{array}{ll} 1 & \text{if } \langle \boldsymbol{w}, \Delta \theta_i(\boldsymbol{y}) \rangle \le 0, \forall \boldsymbol{y} \ne \boldsymbol{y}_i, \boldsymbol{y} \in \mathcal{Y}; \\ 0 & \text{if } \langle \boldsymbol{w}, \Delta \theta_i(\boldsymbol{y}) \rangle > 0, \forall \boldsymbol{y} \ne \boldsymbol{y}_i, \boldsymbol{y} \in \mathcal{Y}. \end{array} \right. \quad (8)$$

Our goal is to find a parameter $\boldsymbol{w}$ that minimizes the empirical error $\hat{R}(\{\boldsymbol{x}_i, \boldsymbol{y}_i\}_{i=1}^n; \boldsymbol{w})$. Considering the computational efficiency, in practice, we use the following convex loss which upper bounds $\ell(\boldsymbol{x}_i, \boldsymbol{y}; \boldsymbol{w})$ to avoid directly minimize the step-function loss:

$$\ell_h(\boldsymbol{x}_i, \boldsymbol{y}; \boldsymbol{w}) = (1 - \langle \boldsymbol{w}, \Delta \theta_i(\boldsymbol{y}) \rangle)_+ \quad (9)$$

where $(\cdot)_+$ is a hinge loss in classification. Correspondingly, we can now define the following empirical hinge risk which upper bounds $\hat{R}(\{\boldsymbol{x}_i, \boldsymbol{y}_i\}_{i=1}^n; \boldsymbol{w})$:

$$\hat{R}_h(\{\boldsymbol{x}_i, \boldsymbol{y}_i\}_{i=1}^n; \boldsymbol{w}) = \frac{1}{n} \sum_{i=1}^n \sum_{\boldsymbol{y} \ne \boldsymbol{y}_i, \boldsymbol{y} \in \mathcal{Y}} \ell_h(\boldsymbol{x}_i, \boldsymbol{y}; \boldsymbol{w}) \quad (10)$$

Accordingly, we can formulate a regularized version of $\hat{R}_h(\{\boldsymbol{x}_i, \boldsymbol{y}_i\}_{i=1}^n; \boldsymbol{w})$ that minimizes an appropriate combination of the empirical error and a regularization term $\Omega(||\boldsymbol{w}||^2)$ to avoid overfitting of the learned model. That is

$$\min_{\boldsymbol{w}} \left\{ \hat{R}_h(\{\boldsymbol{x}_i, \boldsymbol{y}_i\}_{i=1}^n; \boldsymbol{w}) + \lambda \cdot \Omega \left(||\boldsymbol{w}||^2 \right) \right\} \quad (11)$$

where Ω is a strictly monotonically increasing function, and λ is a parameter trading off between the empirical risk and the regularizer. As indicated in [4], such a regularization term can give some smoothness to the obtained function so that the nearby mapped $\theta(\boldsymbol{x}, \boldsymbol{y}), \theta(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}})$ have the similar function value $F(\theta(\boldsymbol{x}, \boldsymbol{y}); \boldsymbol{w}), F(\theta(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}}); \boldsymbol{w})$. Such a local smoothness assumption is intuitive and can relieve the negative influence of the noise training data.

In practice, the above optimization problem can be solved by reducing it to a convex quadratic problem. Similar to what is done in SVMs [4], by introducing a slack variable $\xi_i(\boldsymbol{y})$ for each pair $(\boldsymbol{x}_i, \boldsymbol{y})$, the optimization formulation in (11) can be rewritten as

$$\begin{aligned}\min_{\boldsymbol{w}} \tfrac{1}{2}||\boldsymbol{w}||^2 + \tfrac{\lambda}{n} \cdot \sum_{i=1}^n \sum_{\boldsymbol{y} \ne \boldsymbol{y}_i, \boldsymbol{y} \in \mathcal{Y}} \xi_i(\boldsymbol{y}) \\ s.t. \langle \boldsymbol{w}, \Delta \theta_i(\boldsymbol{y}) \rangle \ge 1 - \xi_i(\boldsymbol{y}), \xi_i(\boldsymbol{y}) \ge 0 \, \boldsymbol{y} \ne \boldsymbol{y}_i, \boldsymbol{y} \in \mathcal{Y}\end{aligned} \quad (12)$$

On introducing Lagrange multipliers $\alpha_i(\boldsymbol{y})$ into the above inequalities and formulating the Lagrangian dual according to Karush-Kuhn-Tucker (KKT) theorem [1], the above problem further reduces to the following convex quadratic problem (QP):

$$\max_{\alpha} \sum_{i, \boldsymbol{y} \ne \boldsymbol{y}_i} \alpha_i(\boldsymbol{y}) - \frac{1}{2} \sum_{i, \boldsymbol{y} \ne \boldsymbol{y}_i} \sum_{j, \tilde{\boldsymbol{y}} \ne \boldsymbol{y}_j} \alpha_i(\boldsymbol{y}) \alpha_j(\tilde{\boldsymbol{y}}) \langle \Delta \theta_i(\boldsymbol{y}), \Delta \theta_j(\tilde{\boldsymbol{y}}) \rangle$$

$$s.t. 0 \le \sum_{\boldsymbol{y} \ne \boldsymbol{y}_i, \boldsymbol{y} \in \mathcal{Y}} \alpha_i(\boldsymbol{y}) \le \frac{\lambda}{n}, \boldsymbol{y} \ne \boldsymbol{y}_i, \boldsymbol{y} \in \mathcal{Y}, 1 \le i \le n \quad (13)$$

and the equality

$$\boldsymbol{w} = \sum_{1 \le i \le n, \boldsymbol{y} \in \mathcal{Y}} \alpha_i(\boldsymbol{y}) \Delta \theta_i(\boldsymbol{y}) \quad (14)$$

Different from those dual variables in the conventional SVMs which only depend on the training data of observation and the associated label pairs $(\boldsymbol{x}_i, \boldsymbol{y}_i), 1 \le i \le n$, the Lagrangian duals in (13) depend on all assignment of labels $\boldsymbol{y}$, which are not limited to the true label of $\boldsymbol{y}_i$. We can iteratively find the active constraints and the associated label variable $\boldsymbol{y}^*$ which most violates the constraints in (9) as $\boldsymbol{y}^* = \arg \max_{\boldsymbol{y} \ne \boldsymbol{y}_i} F(\boldsymbol{x}_i, \boldsymbol{y}; \boldsymbol{w})$ and $\Delta F_i(\boldsymbol{y}^*) < 1$. An active set is maintained for these corresponding active dual variables $\alpha_i(\boldsymbol{y}^*)$, and $\boldsymbol{w}$ is optimized over this set during each iteration using commonly available QP solvers (e.g. SMO [4]).

3. CONNECTION WITH GIBBS RANDOM FIELDS FOR MULTI-LABEL REPRESENTATION

In this section we give an intuitive interpretation of our multi-labeling model through Gibbs Random Fields (GRFs). Detailed mathematical introduction about GRFs can be found in [19]. We can rewrite Eqn. (1) as

$$\begin{aligned}F(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{w}) &= \langle \boldsymbol{w}, \theta(\boldsymbol{x}, \boldsymbol{y}) \rangle \\ &= \sum_{p \in \wp} D_p(y_p; \boldsymbol{x}) + \sum_{(p,q) \in \mathcal{N}} V_{p,q}(y_p, y_q; \boldsymbol{x})\end{aligned} \quad (15)$$
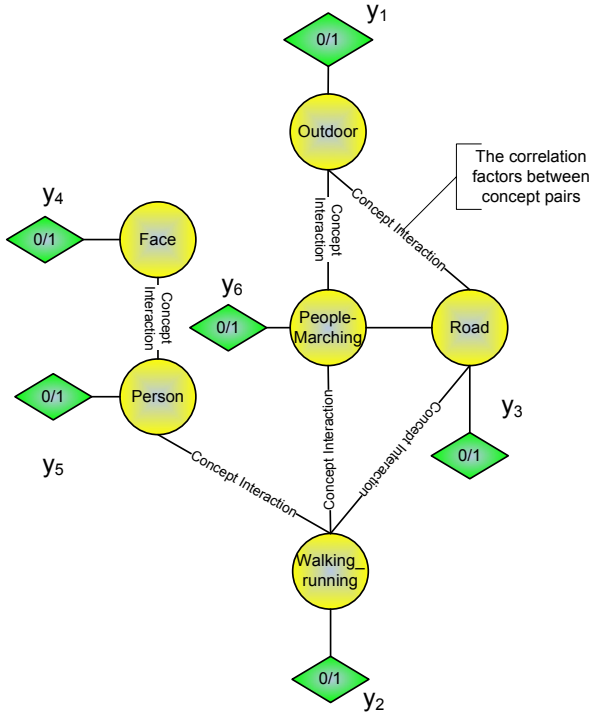
Figure 2: Gibbs Random Fields for a correlative multi-label representation. The edges between concepts indicate the correlation factors $P_{p,q}(y_p, y_q|x)$ between concept pairs.

and

$$D_p(y_p; \boldsymbol{x}) = \sum_{1 \le d \le D, l \in \{+1, -1\}} \boldsymbol{w}_{d,p}^l \theta_{d,p}^l(\boldsymbol{x}, \boldsymbol{y})$$
$$V_{p,q}(y_p, y_q; \boldsymbol{x}) = \sum_{m,n \in \{+1, -1\}} \boldsymbol{w}_{p,q}^{m,n} \theta_{p,q}^{m,n}(\boldsymbol{x}, \boldsymbol{y}) \quad (16)$$

where $\wp = \{i|1 \le i \le K\}$ is a finite index set of the concepts with every $p \in \wp$ representing a video concept, and $\mathcal{N} = \{(p,q)|1 \le p < q \le K\}$ is the set of interacting concept pairs. From the GRFs point of view, $\wp$ is the set of sites of a random field and $\mathcal{N}$ consists of adjacent sites of the concepts. For example, in Figure 2, the corresponding GRF has 6 sites representing "outdoor", "face", "person", "people marching", "road" and "walking running", and these sites are interconnected by the concept interactions, such as (outdoor, people marching), (face, person), (people marching, walking running) etc., which are included in the neighborhood set $\mathcal{N}$ of GRF. In the CML framework, the corresponding $\mathcal{N}$ consists of all pairs of the concepts, i.e., this GRF has a fully connected structure.

Now we can define the energy function for GRF given an example $\boldsymbol{x}$ as

$$H(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{w}) = -F(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{w})$$
$$= -\left\{ \sum_{p \in \wp} D_p(y_p; \boldsymbol{x}) + \sum_{(p,q) \in \mathcal{N}} V_{p,q}(y_p, y_q; \boldsymbol{x}) \right\} \quad (17)$$

and thus we have the probability measure for a particular concept label vector $\boldsymbol{y}$ given $\boldsymbol{x}$ in the form

$$P(y|x, w) = \frac{1}{Z(x, w)} \exp\{-H(y|x, w)\} \quad (18)$$

where $Z(\boldsymbol{x}, \boldsymbol{w}) = \sum_{\boldsymbol{y} \in \mathcal{Y}} \exp\{-H(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{w})\}$ is the partition

function. Such a probability function with an exponential form can express a wide range of probabilities that are strictly positive over the set $\mathcal{Y}$ [19]. It can be easily seen that when inferring the best label vector $\boldsymbol{y}$, maximizing $P(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{w})$ according to the *Maximum A Posteriori Probability* (MAP) criterion is equal to minimizing the energy function $H(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{w})$ or equivalently maximizing $F(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{w})$, which accords with Eqn. (2). Therefore, our CML model is essentially equivalent to the above defined GRF.

Based on this GRF representation for multi-labeling video concepts, the CML model now has a natural probability interpretation. Substitute Eqn. (17) into (18), we have

$$P(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{w}) = \frac{1}{Z(\boldsymbol{x}, \boldsymbol{w})} \prod_{p \in \wp} P(y_p|\boldsymbol{x}) \cdot \prod_{(p,q) \in \mathcal{N}} P_{p,q}(y_p, y_q|\boldsymbol{x}) \quad (19)$$

where

$$P(y_p|\boldsymbol{x}) = \exp\{D_p(y_p; \boldsymbol{x})\}$$
$$P_{p,q}(y_p, y_q|\boldsymbol{x}) = \exp\{V_{p,q}(y_p, y_q; \boldsymbol{x})\}$$

Here $P(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{w})$ has been factored into two types of multipliers. The first type, i.e., $P(y_p|\boldsymbol{x})$, accounts for the probability of a label y_p for the concept p given $\boldsymbol{x}$. These factors indeed model the relations between the concept label and the low-level feature $\boldsymbol{x}$. Note that $P(y_p|\boldsymbol{x})$ only consists of the first type of our constructed features in Eqn. (3), and thus it confirms our claim that the first type of the elements in $\theta(\boldsymbol{x}, \boldsymbol{y})$ serves to capture the connections between x and the individual concept labels. The same discussion can be applied to the second type of the multipliers $P_{p,q}(y_p, y_q|\boldsymbol{x})$. These factors serve to model the correlations between the different concepts, and therefore our constructed features in Eqn. (4) account for the correlations of the concept labels.

The above discussion justifies the proposed model and the corresponding constructed feature vector $\theta(\boldsymbol{x}, \boldsymbol{y})$ for the multi-labeling problem on video semantic annotation. In the next section, we will give further discussion based on this GRF representation.

4. IMPLEMENTATION ISSUES

In this section, we will discuss implementation considerations about CML.

4.1 Interacting concepts

In Section 3, we have revealed the connection between the proposed algorithm and GRFs. As has been discussed, the neighborhood set $\mathcal{N}$ is a collection of the interacting concept pairs, and as for CML, this set contains all possible pairs.

However, in practice, some concept pairs may have rather weak interactions, including both positive and negative ones. For example, the concept pairs (airplane, walking running), (people marching, corporate leader) indeed do not have too many correlations, that is to say, the presence/absence of one concept will not contribute to the presence/absence of another concept (i.e., they occur nearly independently). Based on this observation, we can only involve the strongly interacted concept pairs into the set $\mathcal{N}$, and accordingly the kernel function (5) used in CML becomes

$$\langle \theta(\boldsymbol{x}, \boldsymbol{y}), \theta(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}}) \rangle = \langle \boldsymbol{x}, \tilde{\boldsymbol{x}} \rangle \sum_{1 \le k \le K} \delta [\![y_k = \tilde{y}_k]\!]$$
$$+ \sum_{(p,q) \in \mathcal{N}} \delta [\![y_p = \tilde{y}_p]\!] \delta [\![y_q = \tilde{y}_q]\!] \quad (20)$$

The selection of concept pairs can be manually determined by experts or automatically selected by data-driven approaches. In our algorithm, we adopt an automatic selection process
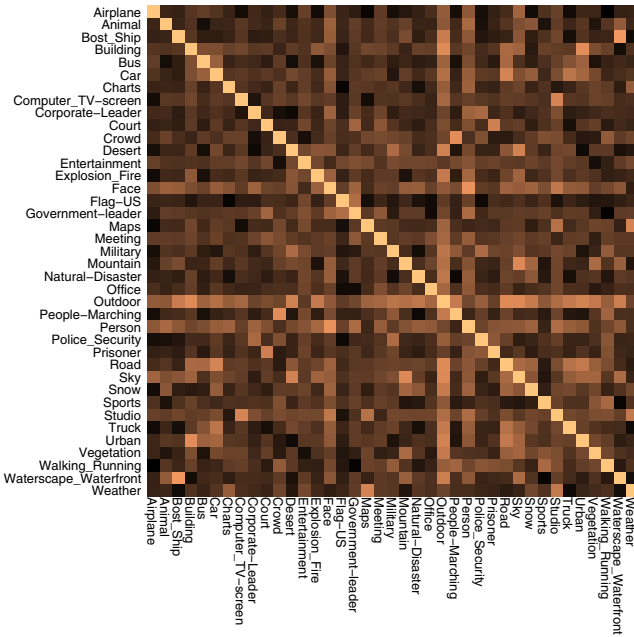
21

Figure 3: The normalized mutual information between each pair of the 39 concepts in the LSCOM-Lite annotations data set. These are computed based on the annotations of the development data set in the experiments (see Section 5).
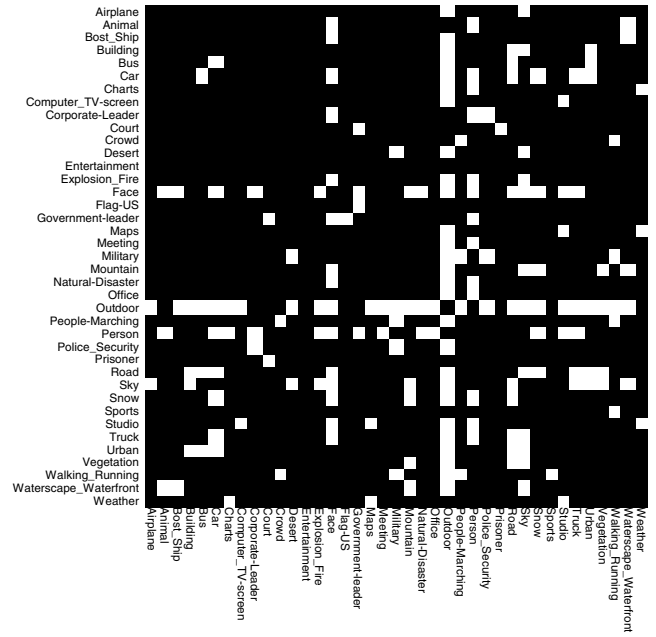

Figure 4: The selected concept pairs according to the computed normalized mutual information. The white blocks indicate the selected concept pairs with significant correlations.

in which the expensive expert labors are not required. First, we use the normalized mutual information [21] to measure the correlations of each concept pair (p, q) as

$$NormMI(p, q) = \frac{MI(p, q)}{\min\{H(p), H(q)\}} \quad (21)$$

where $MI(p, q)$ is the mutual information of the concept p and q, defined by

$$MI(p, q) = \sum_{y_p, y_q} P(y_p, y_q) \log \frac{P(y_p, y_q)}{P(y_p)P(y_q)} \quad (22)$$

and $H(p)$ is the marginal entropy of concept p defined by

$$H(p) = -\sum_{y_p \in \{+1, -1\}} P(y_p) \log P(y_p) \quad (23)$$

Here the label prior probabilities $P(y_p)$ and $P(y_q)$ can be estimated from the labeled ground-truth of the training dataset. According to the information theory [21], the larger the $NormMI(p, q)$ is, the stronger the interaction between concept pair p and q is. Such a normalized measure of concept interrelation has the following advantages:

- It is normalized into the interval [0, 1]: $0 \leq NormMI(p, q) \leq 1$;

- $NormMI(p, q) = 0$ when the concept p and q are statistically independent;

- $NormMI(p, p) = 1$

The above properties are accordant with our intuition about concept correlations, and can be easily proven based on the above definitions. From the above properties, we can find

that the normalized mutual information is scaled into the interval [0, 1] by the minimum concept entropy. With such a scale, the normalized mutual information only considers the concept correlations, which is irrelevant to the distributions of positive and negative examples of the individual concepts. From the normalized mutual information, the concept pairs whose correlations are larger than a threshold are selected. Figure 3 illustrates the normalized mutual information between the 39 concepts in LSCOM-Lite annotation data set. The brighter the grid is, the larger the corresponding normalized mutual information is, and hence the correlation of the concept pair. For example, ("boat ship", "waterscape waterfront"), ("weather", "maps") etc. have larger normalized mutual information. The white dots in Figure 4 represent the selected concept pairs.

4.2 Concept Label Vector Prediction

Once the classification function is obtained, the best predicted concept vector $\boldsymbol{y}^*$ can be obtained from Eqn. (2). The most direct approach is to enumerate all possible label vectors in $\mathcal{Y}$ to find the best one. However, the size of the set $\mathcal{Y}$ will become exponentially large with the increment of the concept number K, and thus the enumeration of all possible concept vectors is practically impossible. For example, when $K = 39$, the size is $2^{39} \approx 5.5 \times 10^{11}$.

Fortunately, from the revealed connection between CML and GRF in Section 4, the prediction of the best concept vector $\boldsymbol{y}^*$ can be performed on the corresponding GRF form. Therefore, many popular approximate inference techniques on GRF can be adopted to predict $\boldsymbol{y}^*$, such as *Annealing Simulation*, *Gibbs Sampling*, etc. Specifically, these approximation techniques will be based on the output optimal dual variables $\alpha_i(\boldsymbol{y})$ in (14). Following the discussion in Section

3, we can give the dual form of the GRF energy function accordingly. Such a dual energy function comes from Eqn. (14). Substituting (14) into (1) and considering the kernel representation (5), we can obtain the following equations:

$$F(\bar{\boldsymbol{x}}, \bar{\boldsymbol{y}}; \boldsymbol{w}) = \left\langle \sum_{1 \le i \le n, y \in \mathcal{Y}} \alpha_i(\boldsymbol{y}) \Delta \theta_i(\boldsymbol{y}), \theta(\bar{\boldsymbol{x}}, \bar{\boldsymbol{y}}) \right\rangle$$
$$= \sum_{p \in \wp} \tilde{D}_p(y_p; \boldsymbol{x}) + \sum_{(p,q) \in \mathcal{N}} \tilde{V}_{p,q}(y_p, y_q; \bar{\boldsymbol{x}}) \quad (24)$$

where

$$\tilde{D}_p(\bar{y}_p; \bar{x}) = \sum_{1 \le i \le n, y \in \mathcal{Y}} \alpha_i(y) k(x_i, \bar{x}) \{ \begin{array}{l} \delta \llbracket y_{ip} = \bar{y}_p \rrbracket - \\ \delta \llbracket y_p = \bar{y}_p \rrbracket \end{array} \}$$
$$\tilde{V}_{p,q}(\bar{y}_p, \bar{y}_q; \bar{x}) =$$
$$\sum_{1 \le i \le n, y \in \mathcal{Y}} \alpha_i(y) \{ \begin{array}{l} \delta \llbracket y_i = \bar{y}_p \rrbracket \delta \llbracket y_{iq} = \bar{y}_q \rrbracket - \\ \delta \llbracket y_p = \bar{y}_p \rrbracket \delta \llbracket y_q = \bar{y}_q \rrbracket \end{array} \}$$
$$(25)$$

And hence the dual energy function is

$$\tilde{H}(\bar{\boldsymbol{y}}|\bar{\boldsymbol{x}}, \boldsymbol{w}) = - \left\{ \begin{array}{l} \sum_{p \in \wp} \tilde{D}_p(\bar{y}_p; \bar{\boldsymbol{x}}) + \\ \sum_{(p,q) \in \mathcal{N}} \tilde{V}_{p,q}(\bar{y}_p, \bar{y}_q; \bar{\boldsymbol{x}}) \end{array} \right\} \quad (26)$$

and the corresponding probability form of GRF can be written as

$$P(\bar{\boldsymbol{y}}|\bar{\boldsymbol{x}}, \boldsymbol{w}) = \frac{1}{\tilde{Z}(\bar{\boldsymbol{x}}, \boldsymbol{w})} \exp \left\{ -\tilde{H}(\bar{\boldsymbol{y}}|\bar{\boldsymbol{x}}, \boldsymbol{w}) \right\} \quad (27)$$

where $\tilde{Z}(\bar{\boldsymbol{x}}, \boldsymbol{w}) = \sum_{y \in \mathcal{Y}} \exp \left\{ -\tilde{H}(y|\bar{\boldsymbol{x}}, \boldsymbol{w}) \right\}$ is the partition function of the dual energy function. With the above dual probabilistic GRF formulation, we use *Iterated Conditional Modes* (ICM) [19] for inference of $\boldsymbol{y}^*$ considering its effectiveness and easy implementation. Other efficient approximation inference techniques (e.g., A*nnealing Simulation*, etc.) can also be directly adopted given the above dual forms.

4.3 Concept Scoring

The output of our algorithm given a sample $\boldsymbol{x}$ is the predicted binary concept label vector. However, for the video retrieval applications, we would like to give each concept of each sample a ranking score for indexing. With these scores, the retrieved video clips can be ranked according to the presence possibility of detecting the concept. Here we give a ranking scoring scheme based on the probability form (Eqn. 27). Given the predicted concept vector $\boldsymbol{y}^*$, the conditional expectation of y_p for the concept p can be computed as

$$E(y_p|\boldsymbol{x}, \boldsymbol{y}^*_{\wp \setminus p}) = P(y_p = +1|\boldsymbol{x}, \boldsymbol{y}^*_{\wp \setminus p})$$
$$- P(y_p = -1|x, \boldsymbol{y}^*_{\wp \setminus p}) \quad (28)$$

where

$$P(y_p|\boldsymbol{x}, \boldsymbol{y}^*_{\wp \setminus p}) = \frac{\exp\{-H(y_p \circ \boldsymbol{y}^*_{\wp \setminus p}|\boldsymbol{x}, \boldsymbol{w})\}}{Z_p}$$
$$= \frac{\exp\{F(\boldsymbol{x}, y_p \circ \boldsymbol{y}^*_{\wp \setminus p}; \boldsymbol{w})\}}{Z_p} \quad (29)$$

and

$$Z_p(\boldsymbol{x}, \boldsymbol{y}^*_{\wp \setminus p}) = \sum_{y_p \in \{+1, -1\}} \exp\{-H(y_p \circ \boldsymbol{y}^*_{\wp \setminus p}|\boldsymbol{x}, \boldsymbol{w})\} \quad (30)$$

is the partition function on the site p. Then we can use this label expectation to rank the video clips for a certain concept.
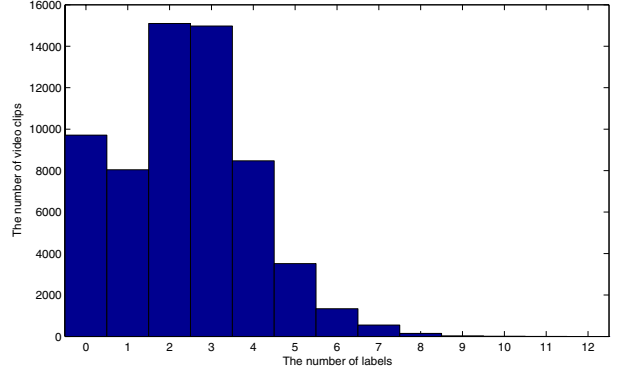

Figure 5: The numbers of labels for the video clips in LSCOM-Lite Annotation data set.

5. EXPERIMENTS

In this section, we evaluate our algorithm on a widely used benchmark video data set and compare it with other state-of-the-art approaches.

5.1 Data Set Description

To evaluate the proposed video annotation algorithm, we conduct the experiments on the benchmark TRECVID 2005 data set [17]. This is one of the most widely used data sets by many groups in the area of multimedia concept modeling[2][3][7]. This data set contains about 170 hours international broadcast news in Arabic, English and Chinese. These news videos are first automatically segmented into $61,901$ subshots. All subshots are then processed to extract several kinds of low-level features, including

1 Block-wise Color Moment in Lab color space;

2 Co-occurrence Texture;

3 Wavelet Texture;

4 Edge Distribution Layout;

and some mid-level features

5 Face - consisting of the face number, face area ratio, the position of the largest face.

For each subshot, 39 concepts are multi-labeled according to LSCOM-Lite annotations [12]. These annotated concepts consist of a wide range of genres, including program category, setting/scene/site, people, object, activity, event, and graphics. Figure 6 illustrates these concepts and their distribution in the data set. Intuitively, many of these concepts have significant semantic correlations between each other. Moreover, these correlations are also proven statistically significant by the normalized mutual information (See Figure 3).

Figure 5 illustrates the multi-labeling nature of the TRECVID data set. As shown, many subshots (71.32%) have more than one label, and some subshots are even labeled with 11 concepts. Such rich multi-labeled subshots in the video data set as well as the significant correlative information between the concepts validate the necessity of exploiting the relationship between the video concepts.
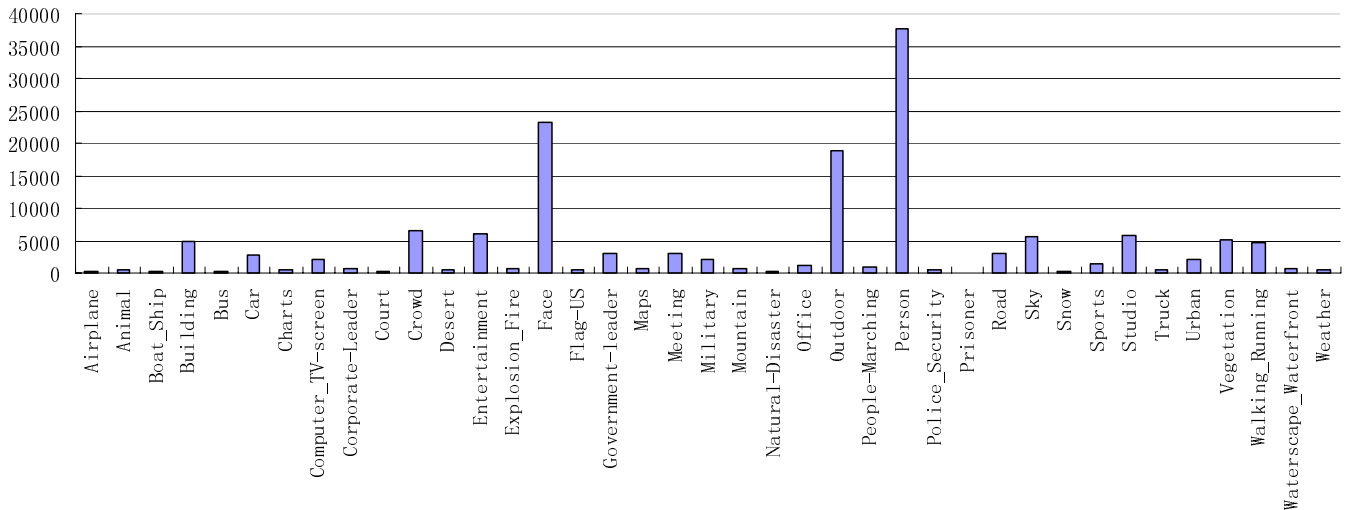
Figure 6: Video Concepts and their distribution in LSCOM-Lite data set

5.2 Experiment Setup

For performance evaluation, we compare our algorithm with two state-of-the-art approaches in first and second paradigms. The first approach, called IndSVM in this section, is the combination of multiple binary encoded SVMs (see the left part of Figure 1.) which are trained independently on each concept; the other approach is developed by adding a contextual fusion level on the detection output of the first approach [5]. In our implementation, we use the SVM for this fusion level. We denote this context-based concept fusion approach as CBCF in this section.

The video data is divided into 3 parts with 65% (40,000 subshots) as training set, 16% (10,000 subshots) as validation set and the remaining 19% (11,901 subshots) as test set. For CBCF, the training set is further split into two parts: one part (32000 subshots) is used for training the individual SVMs in the first detection step, the other part (8000 subshots) is used for training the contextual classifier in the second fusion step. For performance evaluation, we use the official performance metric *Average Precision* (AP) in the TRECVID tasks to evaluate and compare the algorithms on each concept. The AP corresponds to the area under a non-interpolated recall/precision curve and it favors highly ranked relevant subshots. We average the AP over all the 39 concepts to create the mean average precision (MAP), which is the overall evaluation result.

The parameters of the algorithms are determined through a validation process according to their performances on the validation set. For a fair comparison, the results of the all 3 paradigm algorithms reported in this section are the best ones from the chosen parameters. Specifically, two parameters need to be estimated in the proposed CML: the trading-off parameter λ and the Gaussian kernel bandwidth σ of the Gaussian kernel function $\langle \boldsymbol{x}, \tilde{\boldsymbol{x}} \rangle$ in Eqns. (5) and (24). They are respectively selected from sets $\{0.5, 1.0, 10, 100\}$ and $\{0.65, 1.0, 1.5, 2.0\}$ via the validation process. Similarly, the trading-off parameter λ and the Gaussian kernel bandwidth σ in the IndSVM and CBCF are also respectively selected from $\{0.5, 1.0, 10, 100\}$ and $\{0.65, 1.0, 1.5, 2.0\}$, and the best one on the validation set is chosen.

5.3 Experiment Results

In this section, we report experiment results on TRECVID data set. Two different modeling strategies are adopted in the experiments. In the first experiment, all concept pairs are taken into consideration in the model and the kernel function in Eqn. (5) is adopted. We denote this method by CML(I) in our experiment. In the second one, we adopt the strategy described in Section 4.1 and a subset of the concept pairs is applied based on their interacting significance. Accordingly, the kernel function in Eqn. (24) is used, and this approach is denoted by CML(II).

5.3.1 Experiment I

Figure 7 illustrates the performance of CML(I) compared to that of IndSVM (first paradigm) and CBCF (second paradigm). The following observations can be obtained:

- CML(I) obtains about 15.4% and 12.2% relative improvements on MAP compared to IndSVM and CBCF. Compared to the improvement of CBCF (2%) relative to the baseline IndSVM, Such an improvement is significant.

- CML(I) performs the best on 28 of the all 39 concepts. Some of the improvements are significant, such as "office" (477% better than InidSVM and 260% better than CBCF), "people-marching" (68% better than IndSVM and 160% better than CBCF), "walking running" (55% better than IndSVM and 48% better than CBCF).

- CML(I) deteriorates on some concepts compared to IndSVM and CBCF. For example, it has 12% and 14% deterioration on "snow" respectively and 11% and 17% deterioration on "bus" respectively. As discussed in Section 4.1, the performance deterioration is due to insignificant concept relations. Next subsection will present CML(II), which solves this deterioration problem and obtains a more consistent and robust performance improvement.
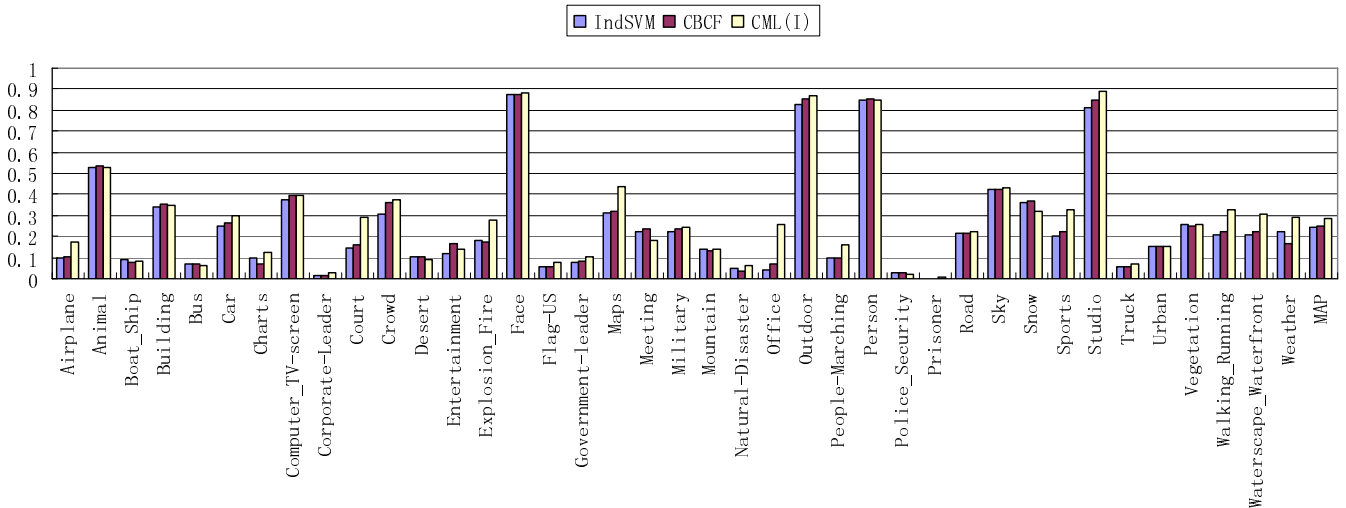
Figure 7: The performance comparison of IndSVM, CBCF and CML(I).

5.3.2 Experiment II

Following the proposed approach in Section 4.1, the deterioration problem can be solved by removing concept pairs with insignificant correlations.

Figure 3 illustrates the normalized mutual entropy between all concepts. They are computed on the development set which includes training set and validation set, but does NOT include the test set. The average normalized mutual information entropy is $Avg_{EN} = 0.02$. An important aspect of a good algorithm is if its parameters can be determined automatically. Following such a principle, the threshold Th_{EN} is automatically determined to be $Th_{EN} = 2Avg_{EN}$ such that any concept pairs whose normalized mutual entropy less than Th_{EN} are removed. Figure 4 shows these selected concept pairs. As we can see, these preserved concept pairs either have intuitive semantic correlations e.g. "waterscape waterfront" and "boat ship" or statistically tend to co-occur in the news broadcast videos, e.g. "maps" and "weather" in weather forecast video subshots.

Figure 8 illustrates the performance of CML(II) with these selected concept pairs compared to IndSVM, CBCF and CML(I). We can find

- CML(II) has the best overall performance compared to the other algorithms. It outperforms IndSVM, CBCF and CML(I) by 17%, 14% and 2%, respectively.

- Furthermore, CML(II) has a more consistent and robust performance improvement over all 39 concepts compared to IndSVM and CBCF. For example, on "bus" and "snow", CML(I) gave worse performance than IndSVM and CBCF. In the contrary, CML(II) gains about 71% and 3% improvement compared to IndSVM and 58% and 1% improvement compared to CBCF with no deterioration.

In summary, CML(II) is the best approach because its best overall MAP improvement as well as its consistent and robust performance on the diverse 39 concepts.

Finally, we give an empirical comparison of computational cost between the proposed CML and the other two state-of-the-art algorithms (IndSVM and CBCF). In fact, under the different parameter settings, the computational cost is different largely. But in general, as for IndSVM and CBCF, the models of each concept are independent without coupled with each other, so they can be trained in parallel. Therefore the computing time needed is much less than CML in which the modeling of the whole concept set is conducted in a coupled manner and is unable to be operated in parallel. In our experiment, the speed of CML is about 25 times slower than IndSVM and CBCF. Thus how to accelerate the computation speed of CML will be the focus of our future work.

6. CONCLUSIONS AND FUTURE WORKS

In this paper, we proposed a correlative multi-labeling (CML) approach to exploit how the concept correlations help infer the video semantic concepts. Different from the first and second paradigms, where they suffer from insufficient modeling of concept correlations, the proposed approach is able to simultaneously model both the individual concept and the conceptual correlations in an integrated framework. In addition, CML is highly efficient in utilizing the data set. Experiments on the widely used benchmark TRECVID data set demonstrated that CML is superior to state-of-the-art approaches in the first and second paradigms, in both overall performance and the consistency of performance on diverse concepts.

We will continue our future works in two directions. First, we will study how the performance changes with the increment of video concept number, and if the algorithm can get more improvement gain by exploiting a large number of concepts. Second, we will also apply the proposed algorithm to other applications, such image annotation, text categorization in which there exists a large number of correlative concepts.

7. REFERENCES

[1] S. Boyd and L. Vandenberghe. *Convex Optimization.* Cambridge University Press, 2004.
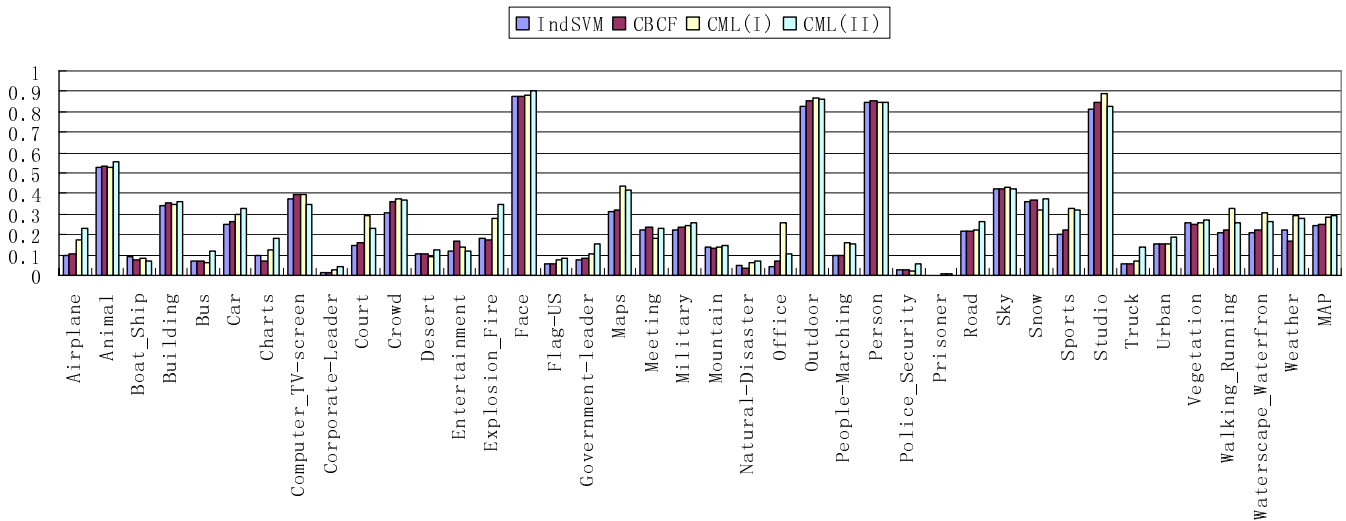
Figure 8: The performance comparison of IndSVM, CBCF and CML(II).

[2] M. Campbell and et al. Ibm research trecvid-2006 video retrieval system. In *TREC Video Retrieval Evaluation (TRECVID) Proceedings*, 2006.

[3] S.-F. Chang and et al. Columbia university trecvid-2006 video search and high-level feature extraction. In *TREC Video Retrieval Evaluation (TRECVID) Proceedings*, 2006.

[4] N. Cristianini and J. Shawe-Taylor. *An introduction to support vector machines and other kernel-based learning methods*. Cambridge University, 2000.

[5] S. Godbole and S. Sarawagi. Discriminative methods for multi-labeled classification. In *PAKDD*, 2004.

[6] A. Hauptmann, M.-Y. Chen, and M. Christel. Confounded expectations: Informedia at TRECVID 2004. In *TREC Video Retrieval Evaluation Online Proceedings*, 2004.

[7] A. G. Hauptmann and et al. Multi-lingual broadcast news retrieval. In *TREC Video Retrieval Evaluation (TRECVID) Proceedings*, 2006.

[8] W. Jiang, S.-F. Chang, and A. Loui. Active concept-based concept fusion with partial user labels. In *Proceedings of IEEE International Conference on Image Processing*, 2006.

[9] D. Marr. *Vision*. W.H.Freeman and Company, 1982.

[10] M. Naphade, I. Kozintsev, and T. Huang. Factor graph framework for semantic video indexing. *IEEE Trans. on CSVT*, 12(1), Jan. 2002.

[11] M. R. Naphade. Statistical techniques in video data management. In *IEEE Workshop on Multimedia Signal Processing*, 2002.

[12] 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.

[13] K. Nigam, J. Lafferty, and A. McCallum. Using maximum entropy for text classification. In *IJCAI-99 Workshop on Machine Learning for Information Filtering*, pages 61–67, 1999.

[14] X. Shen, M. Boutell, J. Luo, and C. Brown. Multi-label machine learning and its application to semantic scene classification. In *International Symposium on Electronic Imaging*, 2004.

[15] J. R. Smith and M. Naphade. Multimedia semantic indexing using model vectors. In *Proceeding of IEEE International Conferences on Multimedia and Expo*, 2003.

[16] C. Snoek and et al. The challenge problem for automated detection of 101 semantic concepts in multimedia. In *Proceedings of the ACM International Conference on Multimedia*, pages 421–430, Santa Barbara, USA, October 2006.

[17] TRECVID. http://www-nlpir.nist.gov/projects/trecvid/.

[18] I. Tsochantaridis, T. Hofmann, T. Joachims, and Y. Altun. Support vector machine learning for intedependent and structured output spaces. In *Proc. of Internatial Conference on ICML*, 2004.

[19] G. Winkler. *Image analysis, random fields and dynamic Monte Carlo methods: A mathematical introduction*. Springer-Verlag, Berlin, Heidelberg, 1995.

[20] Y. Wu, B. L. Tseng, and J. R. Smith. Ontology-based multi-classification learning for video concept detection. In *Proceeding of IEEE International Conferences on Multimedia and Expo*, 2004.

[21] Y. Y. Yao. *Entropy measures, maximum entropy principle, and emerging applications*, chapter Information-theoretic measures for knowledge discovery and data mining, pages 115–136. Springer, 2003.