

A Unifying Multi-Label Temporal Kernel Machine with Its Application to Video Annotation

Guo-Jun Qi
University of Science and Technology of China
Xian-Sheng Hua
Microsoft Research Asia
Yong Rui
Microsoft China R&D Group
and
Hong-Jiang Zhang
Microsoft Advanced Technology Center

Automatic video annotation is an important ingredient for semantic-level video browsing, search and navigation. Much attention has been paid to this topic in recent years. These researches have evolved through two paradigms. In the first paradigm, each concept is individually annotated by a pre-trained binary classifier. However, this method ignores the rich information between the video concepts and only achieves limited success. Evolved from the first paradigm, the methods in the second paradigm add an extra step on the top of the first individual classifiers to fuse the multiple detections of the concepts. However due to the unreliable classifiers in the first step, the performance of these methods can be degraded by the errors incurred in the first step. In this paper, another paradigm of the video annotation method is proposed to address the above problems. It simultaneously annotates the concepts as well as model correlations between them in one step by the proposed *Correlative Multi-Label* (CML) method. Furthermore since the video clips are composed by temporal-ordered frame sequences, we extend the proposed method to exploit the rich temporal information in the videos. Specifically, a temporal-kernel is incorporated into the CML method based on the discriminative information between *Hidden Markov Models* (HMM) that are learned from the video clips. We compare the performance between the proposed approach and the state-of-the-art approaches in the first and second paradigms on the widely used TRECVID data set. As to be shown, superior performance from the proposed method can be gained.

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

Additional Key Words and Phrases: Video Annotation, Multi-Labeling, Concept Correlation, Temporal Kernel

Author's address: G.-J. Qi, Department of Automation, University of Science and Technology of China, 96 Jinzhai Road, Hefei, Anhui 230027, China; email: qgj@mail.ustc.edu.cn


X.-S. Hua, Microsoft Research Asia, 49 Zhichun Road, 100080 Beijing, China; email:xshua@microsoft.com

Y. Rui, Microsoft China R&D Group, 49 Zhichun Road, 100080 Beijing, China; email:yongrui@microsoft.com

H.-J. Zhang, Microsoft Advanced Technology Center, 49 Zhichun Road, 100080 Beijing, China; email:hjzhang@microsoft.com

Permission to make digital/hard copy of all or part of this material without fee for personal or classroom use provided that the copies are not made or distributed for profit or commercial advantage, the ACM copyright/server notice, the title of the publication, and its date appear, and notice is given that copying is by permission of the ACM, Inc. To copy otherwise, to republish, to post on servers, or to redistribute to lists requires prior specific permission and/or a fee.

© 20XX ACM 0000-0000/20XX/0000-0001 \$5.00



Outdoor	T	T	T	T	T	T	T
Face	T	T	T	T	T	T	T
Person	T	T	T	T	T	T	T
People-Marching	F	F	F	T	T	F	F
Road	T	T	T	T	T	T	T
Walking_running	T	T	T	T	T	T	T

Fig. 1. Some multi-labeled examples from TRECVID dataset.

1. INTRODUCTION

With explosive emergence of considerable videos on the Internet (e.g., Youtube, VideoEgg, Yahoo! Video, and many video clips on personal homepages and blogs etc.), effective indexing and searching these video corpus becomes more and more attractive to users. As a basic technique in video index and search, semantic-level video annotation has been an important research topic in the multimedia research community [Naphade 2002][Snoek et al. 2006]. It aims at annotating video clips with a set of the concepts of interest, including 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.)[Naphade et al. 2005][Snoek et al. 2006]. In this paper, we are concerned with a multi-label video annotation process where a video clip can be annotated by multiple labels at the same time. Figure 1 illustrates some keyframes of the multi-labeled video clips. For example, a video clip can be classified as “person”, “walking_running” and “road” simultaneously. In contrast to the multi-label problem, multi-class annotation only assigns one concept to each video clip. In most real-world video annotations, such as TRECVID annotations and the users’ tags on many video-sharing website, the video clips are often multi-labeled by a set of the concepts rather than only a single one. Since it involves nonexclusive classification of multiple concepts, multi-label annotation is much more complex than multi-class annotation. We will focus on multi-label video annotation in this paper.

1.1 Video Annotation with Multiple Labels

Multi-label video annotation has evolved through two paradigms: individual concept detection and annotation, and *Context Based Conceptual Fusion* (CBCF) [Jiang et al. 2006] annotation. In this paper, we propose the third paradigm: the unifying multi-label annotation. We next review these three paradigms.

1.1.1 Paradigm I: Individual Concept Annotation. The annotation methods in the first paradigm are individual concept detectors, i.e., they annotate the video concepts individually and independently. They neglect the rich correlations between the video concepts. In more detail, these methods only translate the multi-label annotations into some independent concept detectors which individually assign presence/absence labels into each sam-

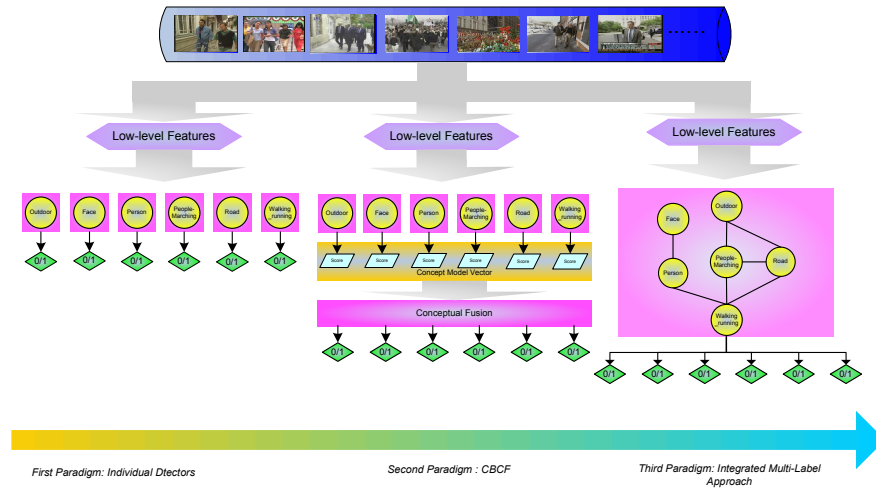


Fig. 2. The multi-label video annotation methods in three paradigms. From leftmost to the rightmost, they are the individual SVM, CBCF and our proposed CML.

ple. Most classical detectors can be categorized into this paradigm. For example, Support Vector Machine (SVM) [Cristianini and Shawe-Taylor 2000] with one-against-the-other strategy attempts to learn a set of detectors, each of which independently models the presence/absence of a certain concept. Other examples of this paradigm contain Maximum Entropy Models (MEM) [Nigam et al. 1999], Manifold Ranking (MR) [Tang et al. 2007] etc. In Figure 2, we give an illustration of this paradigm in the leftmost flowchart. As depicted, a set of individual SVMs is learned for video concept annotation. In brief, the core of this paradigm is to formulate the video annotation as a collection of independent binary classifiers.

However in many real-world problems, video concepts often exist correlatively with each other, rather than appear in isolation. So the individual annotation only achieves limited success. For example, the presence of “crowd” often occurs together with the presence of “people” while “boat ship” and “truck” commonly do not co-occur. On the other hand, compared to simple concepts which can be directly modeled from low-level features, some complex concepts e.g., “people marching”, are really difficult to be individually modeled due to the semantic gap between these concepts and low-level features. Instead, these complex concepts can be better inferred based on the label 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. Therefore, it will be very helpful to exploit the label correlations when annotating the multiple concepts together.

1.1.2 Paradigm II: Context Based Conceptual Fusion Annotation. As a step towards more advanced video annotation, the second paradigm is built atop the individual concept detectors. It attempts to refine the detection results of the binary concept detectors with a Context Based Concept Fusion (CBCF) strategy. Many algorithms can be categorized into this paradigm. For example, [Wu et al. 2004] uses 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 and Naphade 2003] present a two-step Discriminative Model Fusion (DMF) approach 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 2 shows such a second-paradigm approach. Alternative fusion strategy can also be used, e.g., [Hauptmann et al. 2004] propose to use Logistic Regression (LR) to fuse the individual detections. [Jiang et al. 2006] use a CBCF-based active learning method. Users are 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. [Naphade et al. 2002] propose a probabilistic Bayesian Multi-net approach to explicitly model the relationship between the multiple concepts through a factor graph which is built upon the underlying video ontology semantics. [Zha et al. 2007] propose to leverage the pairwise concurrent relations to refine the video detection output by individual classifiers of the concepts.

Intuitively it is reasonable to leverage the context-based conceptual information to improve the accuracy of the concept detectors. However there also exist some experiments to show that the CBCF methods do not have a consistent improvement over the individual detectors. Its overall performance can even be worse than the binary-based detectors. For example, in [Hauptmann et al. 2004] 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 atop 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 methods do not follow the *principle of Least-Commitment* espoused by D. Marr [Marr 1982], 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.

1.1.3 Paradigm III: Unifying Multi-label Annotation. In this paper, we will propose the third paradigm of video annotation to address the problem faced in the first and second paradigms. This new paradigm will simultaneously model both the individual concepts and their correlations in a unifying formulation, and the *principle of Least-Commitment* will be obeyed. The rightmost flowchart of Figure 2 illustrates the proposed *Correlative Multi-Label (CML)* method. As we can see, this method has the following advantages compared to the second CBCF paradigm:

- (1) The approach follows the *Principle of Least-Commitment* [Marr 1982]. 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. In contrast, the third paradigm addresses the correlation issue at the root in a single step.

1.2 Video Annotation with Temporal-ordered sequences

Besides the above multi-label problem, it is also an important issue to leverage the rich temporal information in the video sequences to boost the video annotation, especially for annotating the event-related concepts, such as “airplane-flying”, “riot”, “people-marching” etc.

There already exist some research works which attempt to utilize temporal information for video annotation. These researches have evolved through two research categories. In the first category, some statistical models of feature dynamics are used to represent and detect video semantics. For example, [Xie and Chang 2002] proposed to detect and segment the “play” and “break” events in soccer videos by learning the dynamics of the color and motion features with *Hidden Markov Model* (HMM). This method is only based on low-level feature dynamics to construct a generative model and ignores the other intuitive semantic components, such as visual concept interactions [Ebadollahi et al. 2006]. For example, while detecting “airplane-flying”, it is helpful to detect whether “sky”, “airplane” occurs.

The second research category utilizes the concept interactions to detect the video events. For example, [Ebadollahi et al. 2006] propose to leverage stochastic temporal processes in the concept space to model the video events. This method aims at learning the dynamics of concurrent concepts from exemplars of an event in a pure data-driven fashion. However, these concurrent concepts are obtained from the output of some pre-learned concept detectors. These detectors are often not robust enough to give reliable concept predictions. Therefore, the errors in the first step of concept predictions can propagate to the second step of learning the concept dynamics of the video events. It also violates the *principle of Least Commitment* [Marr 1982] so that the errors incurred in the individual concept detectors cannot be corrected in the second step of learning concept dynamics. The same problem incurs in [Wang et al. 2006] as well. It first pre-trains some mid-level keyword detectors, based on which a Conditional Random Fields (CRF) [Lafferty et al. 2001] are used to capture the interactions between the noisy predictions of these keyword detectors.

To address the above problem, we will introduce a temporal kernel under the proposed correlative multi-label formulation. It can leverage the concept interactions as well as low-level feature dynamics to boost the video event detections. Specifically, it constructs a temporal kernel by revealing the discriminative information between the statistical models that are learned from the video sequences. As to be seen later, it avoids the two-step method in which the noisy outputs of the individual concept detectors is propagated into the second conceptual dynamics. Instead, the concept interactions and low-level feature dynamics are captured in a unifying framework, thus the *principle of Least Commitment* are obeyed. Furthermore, the proposed temporal kernel can be naturally incorporated to the proposed multi-label kernel without any extra complexity of the algorithm.

The rest of the paper is organized as follows. In Section 2, we give a detailed description of the proposed *Correlative Multi-Label* (CML) method, including the classification model, the learning strategy. Furthermore we will explore the connection between the proposed approach and *Gibbs Random Fields* (GRFs) [Winkler 1995], 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 3 details the temporal kernel for video annotation. This kernel can be naturally incorporated into CML kernel to form a *Correlative Multi-Label Temporal* (CMLT) Kernel, which captures the high-level concept interactions and low-level feature dynamics in a unifying kernel machine. Finally, in Section 4, we will report experiments on the benchmark TRECVID data and show that the proposed approach has superior performance over the state-of-the-art algorithms in both first and second paradigms.

2. CORRELATIVE MULTI-LABEL VIDEO ANNOTATION

In this section, we will introduce the 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. In section 2.3, we will give a probabilistic interpretation of the CML model based on Gibbs random fields.

2.1 Multi-Label Classification Function

Before we move further, we first define some notations. Let $\mathbf{x} = (x_1, x_2, \dots, x_D)^T \in \mathcal{X}$ denote the input pattern representing feature vectors extracted from video clips; Let $\mathbf{y} \in \mathcal{Y} = \{+1, -1\}^K$ denote the K dimensional concept label vector of an example, where each entry $y_i \in \{+1, -1\}$ of \mathbf{y} indicates the membership of this example in the i th 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

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

where $\theta(\mathbf{x}, \mathbf{y})$ 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); \mathbf{w} is the linear combination weight vector. With such a discriminative function, for an input pattern \mathbf{x} , the label vector \mathbf{y}^* can be predicted by maximizing over the argument \mathbf{y} as

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

As to be presented in the next section, such a discriminative function can be intuitively interpreted in the Gibbs random fields (GRFs) [Winkler 1995] framework when considering the defined feature vector $\theta(\mathbf{x}, \mathbf{y})$. The constructed feature $\theta(\mathbf{x}, \mathbf{y})$ 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:

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

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 $\theta(\mathbf{x}, \mathbf{y})$ serve to model the connection between the low level feature \mathbf{x} and the labels $y_k (1 \leq k \leq K)$ 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 $\theta(\mathbf{x}, \mathbf{y})$ is required to investigate the correlations between the different concepts.

Type II. The elements for concept correlations:

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

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.

By concatenating the above two types of elements together, we can obtain the feature vector $\theta(\mathbf{x}, \mathbf{y})$. It is not difficult to see that the dimension of vector $\theta(\mathbf{x}, \mathbf{y})$ is $2KD + 4C_K^2 = 2K(D + K - 1)$. When K and D are large, the dimension of $\theta(\mathbf{x}, \mathbf{y})$ will be extraordinary high. For example, if $K = 39$ and $D = 200$, $\theta(\mathbf{x}, \mathbf{y})$ 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, $\theta(\mathbf{x}, \mathbf{y})$ and $\theta(\tilde{\mathbf{x}}, \tilde{\mathbf{y}})$, can be represented in a very compact form as

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

where $\langle \mathbf{x}, \tilde{\mathbf{x}} \rangle$ is the dot product over the low-level feature vector x and \tilde{x} . We call this kernel the *Correlative Multi-Label (CML) Kernel* and the corresponding video annotation method *Correlative Multi-Label Video Annotation* in this paper. It is worth noting that, any other kernel function $K(\mathbf{x}, \tilde{\mathbf{x}})$ (such as Gaussian Kernel, Polynomial Kernel) can be substituted for $\langle \mathbf{x}, \tilde{\mathbf{x}} \rangle$ 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 $\theta(\mathbf{x}, \mathbf{y})$.

2.2 Learning the Classification Function

In this section, we will introduce how to train the classification model (1) with the presented kernel (5). The procedure follows a similar derivation as in the conventional SVM (details about SVM can be found in [Cristianini and Shawe-Taylor 2000]) and in particular one of

its variants for the structural output spaces [Tsochantaridis et al. 2004]. Given an example \mathbf{x}_i and its label vector \mathbf{y}_i from the training set $\{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^n$, according to Eqn. (1) and (2), a misclassification occurs when we have

$$\Delta F_i(\mathbf{y}) \triangleq F(\mathbf{x}_i, \mathbf{y}_i) - F(\mathbf{x}_i, \mathbf{y}) = \langle \mathbf{w}, \Delta\theta_i(\mathbf{y}) \rangle \leq 0, \forall \mathbf{y} \neq \mathbf{y}_i, \mathbf{y} \in \mathcal{Y} \quad (6)$$

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

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

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

$$\ell(\mathbf{x}_i, \mathbf{y}; \mathbf{w}) = \begin{cases} 1 & \text{if } \langle \mathbf{w}, \Delta\theta_i(\mathbf{y}) \rangle \leq 0, \forall \mathbf{y} \neq \mathbf{y}_i, \mathbf{y} \in \mathcal{Y}; \\ 0 & \text{if } \langle \mathbf{w}, \Delta\theta_i(\mathbf{y}) \rangle > 0, \forall \mathbf{y} \neq \mathbf{y}_i, \mathbf{y} \in \mathcal{Y}. \end{cases} \quad (8)$$

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

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

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

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

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

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

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 [Cristianini and Shawe-Taylor 2000], such a regularization term can give some smoothness to the obtained function so that the nearby mapped $\theta(\mathbf{x}, \mathbf{y}), \theta(\tilde{\mathbf{x}}, \tilde{\mathbf{y}})$ have the similar function value $F(\theta(\mathbf{x}, \mathbf{y}); \mathbf{w}), F(\theta(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}); \mathbf{w})$. 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 [Cristianini and Shawe-Taylor 2000], by introducing a slack variable $\xi_i(\mathbf{y})$ for each pair $(\mathbf{x}_i, \mathbf{y})$, the optimization formulation in (11) can be rewritten as

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

On introducing Lagrange multipliers $\alpha_i(\mathbf{y})$ into the above inequalities and formulating the Lagrangian dual according to Karush-Kuhn-Tucker (KKT) theorem [Boyd and Vandenberghe 2004], the above problem further reduces to the following convex quadratic

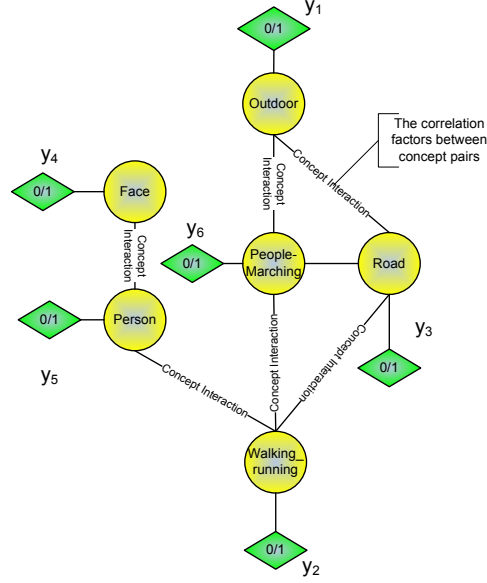


Fig. 3. Gibbs Random Fields for a correlative multi-label representation. The edges between concepts indicate the correlation factors $P_{p,q}(y_p, y_q|x)$ between concept pairs.

problem (QP):

$$\begin{aligned} \max_{\alpha} \sum_{i, \mathbf{y} \neq \mathbf{y}_i} \alpha_i(\mathbf{y}) - \frac{1}{2} \sum_{i, \mathbf{y} \neq \mathbf{y}_i} \sum_{j, \tilde{\mathbf{y}} \neq \mathbf{y}_j} \alpha_i(\mathbf{y}) \alpha_j(\tilde{\mathbf{y}}) \langle \Delta\theta_i(\mathbf{y}), \Delta\theta_j(\tilde{\mathbf{y}}) \rangle \\ s.t. 0 \leq \sum_{\mathbf{y} \neq \mathbf{y}_i, \mathbf{y} \in \mathcal{Y}} \alpha_i(\mathbf{y}) \leq \frac{\lambda}{n}, \mathbf{y} \neq \mathbf{y}_i, \mathbf{y} \in \mathcal{Y}, 1 \leq i \leq n \end{aligned} \quad (13)$$

and the equality

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

Different from those dual variables in the conventional SVMs which only depend on the training data of observation and the associated label pairs $(\mathbf{x}_i, \mathbf{y}_i), 1 \leq i \leq n$, the Lagrangian duals in (13) depend on all assignment of labels \mathbf{y} , which are not limited to the true label of \mathbf{y}_i . We can iteratively find the active constraints and the associated label variable \mathbf{y}^* which most violates the constraints in (9) as $\mathbf{y}^* = \arg \max_{\mathbf{y} \neq \mathbf{y}_i} F(\mathbf{x}_i, \mathbf{y}; \mathbf{w})$ and $\Delta F_i(\mathbf{y}^*) < 1$. An active set is maintained for these corresponding active dual variables $\alpha_i(\mathbf{y}^*)$, and \mathbf{w} is optimized over this set during each iteration using commonly available QP solvers (e.g. SMO [Cristianini and Shawe-Taylor 2000]).

2.3 A Justification - 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 [Winkler 1995]. We can rewrite Eqn. (1) as

$$F(\mathbf{x}, \mathbf{y}; \mathbf{w}) = \langle \mathbf{w}, \theta(\mathbf{x}, \mathbf{y}) \rangle = \sum_{p \in \varnothing} D_p(y_p; \mathbf{x}) + \sum_{(p,q) \in \mathcal{N}} V_{p,q}(y_p, y_q; \mathbf{x}) \quad (15)$$

and

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

where $\wp = \{i | 1 \leq i \leq K\}$ is a finite index set of the concepts with every $p \in \wp$ representing a video concept, and $\mathcal{N} = \{(p, q) | 1 \leq p < q \leq K\}$ 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 3, 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 \mathbf{x} as

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

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

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

where $Z(\mathbf{x}, \mathbf{w}) = \sum_{\mathbf{y} \in \mathcal{Y}} \exp \{-H(\mathbf{y} | \mathbf{x}, \mathbf{w})\}$ 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} [Winkler 1995]. It can be easily seen that when inferring the best label vector \mathbf{y} , maximizing $P(\mathbf{y} | \mathbf{x}, \mathbf{w})$ according to the *Maximum A Posteriori* (MAP) criterion is equal to minimizing the energy function $H(\mathbf{y} | \mathbf{x}, \mathbf{w})$ or equivalently maximizing $F(\mathbf{x}, \mathbf{y}; \mathbf{w})$, which accords with Eqn. (2). Therefore, the 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

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

where

$$\begin{aligned} P(y_p | \mathbf{x}) &= \exp\{D_p(y_p; \mathbf{x})\} \\ P_{p,q}(y_p, y_q | \mathbf{x}) &= \exp\{V_{p,q}(y_p, y_q; \mathbf{x})\} \end{aligned}$$

Here $P(\mathbf{y} | \mathbf{x}, \mathbf{w})$ has been factored into two types of multipliers. The first type, i.e., $P(y_p | \mathbf{x})$, accounts for the probability of a label y_p for the concept p given \mathbf{x} . These factors indeed model the relations between the concept label and the low-level feature \mathbf{x} . Note that $P(y_p | \mathbf{x})$ 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 $\theta(\mathbf{x}, \mathbf{y})$ serves to capture the connections between \mathbf{x} and the individual concept labels. The same discussion can be applied to the second type of the multipliers $P_{p,q}(y_p, y_q | \mathbf{x})$. 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 $\theta(\mathbf{x}, \mathbf{y})$ for the multi-labeling problem on video semantic annotation. In the next section, we will give some further discussions based on this GRF representation.

2.3.1 Concept Label Vector Prediction. Once the classification function is obtained, the best predicted concept vector \mathbf{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 \mathbf{y}^* can be performed on the corresponding GRF form. Therefore, many popular approximate inference techniques on GRF can be adopted to predict \mathbf{y}^* , such as *Annealing Simulation*, *Gibbs Sampling*, etc. Specifically, these approximation techniques will be based on the output optimal dual variables $\alpha_i(\mathbf{y})$ in (14). Following the above discussion about GRF representation, 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:

$$\begin{aligned} F(\bar{\mathbf{x}}, \bar{\mathbf{y}}; \mathbf{w}) &= \left\langle \sum_{1 \leq i \leq n, \mathbf{y} \in \mathcal{Y}} \alpha_i(\mathbf{y}) \Delta \theta_i(\mathbf{y}), \theta(\bar{\mathbf{x}}, \bar{\mathbf{y}}) \right\rangle \\ &= \sum_{p \in \wp} \tilde{D}_p(\bar{y}_p; \bar{\mathbf{x}}) + \sum_{(p,q) \in \mathcal{N}} \tilde{V}_{p,q}(\bar{y}_p, \bar{y}_q; \bar{\mathbf{x}}) \end{aligned} \quad (20)$$

where

$$\begin{aligned} \tilde{D}_p(\bar{y}_p; \bar{\mathbf{x}}) &= \sum_{1 \leq i \leq n, \mathbf{y} \in \mathcal{Y}} \alpha_i(\mathbf{y}) k(x_i, \bar{\mathbf{x}}) \{ \delta \llbracket y_{ip} = \bar{y}_p \rrbracket - \delta \llbracket y_p = \bar{y}_p \rrbracket \} \\ \tilde{V}_{p,q}(\bar{y}_p, \bar{y}_q; \bar{\mathbf{x}}) &= \sum_{1 \leq i \leq n, \mathbf{y} \in \mathcal{Y}} \alpha_i(\mathbf{y}) \{ \delta \llbracket y_i = \bar{y}_p \rrbracket \delta \llbracket y_{iq} = \bar{y}_q \rrbracket - \delta \llbracket y_p = \bar{y}_p \rrbracket \delta \llbracket y_q = \bar{y}_q \rrbracket \} \end{aligned} \quad (21)$$

And hence the dual energy function is

$$\tilde{H}(\bar{\mathbf{y}}|\bar{\mathbf{x}}, \mathbf{w}) = - \left\{ \sum_{p \in \wp} \tilde{D}_p(\bar{y}_p; \bar{\mathbf{x}}) + \sum_{(p,q) \in \mathcal{N}} \tilde{V}_{p,q}(\bar{y}_p, \bar{y}_q; \bar{\mathbf{x}}) \right\} \quad (22)$$

and the corresponding probability form of GRF can be written as

$$P(\bar{\mathbf{y}}|\bar{\mathbf{x}}, \mathbf{w}) = \frac{1}{\tilde{Z}(\bar{\mathbf{x}}, \mathbf{w})} \exp \left\{ -\tilde{H}(\bar{\mathbf{y}}|\bar{\mathbf{x}}, \mathbf{w}) \right\} \quad (23)$$

where $\tilde{Z}(\bar{\mathbf{x}}, \mathbf{w}) = \sum_{\mathbf{y} \in \mathcal{Y}} \exp \left\{ -\tilde{H}(\mathbf{y}|\bar{\mathbf{x}}, \mathbf{w}) \right\}$ is the partition function of the dual energy function. With the above dual probabilistic GRF formulation, we use *Iterated Conditional Modes* (ICM) [Winkler 1995] for inference of \mathbf{y}^* considering its effectiveness and easy implementation. Other efficient approximation inference techniques (e.g., *Annealing Simulation*, etc.) can also be directly adopted given the above dual forms.

2.3.2 Concept Scoring. The output of our algorithm given a sample \mathbf{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. 23). Given the predicted concept vector \mathbf{y}^* , the conditional expectation of y_p for the concept p

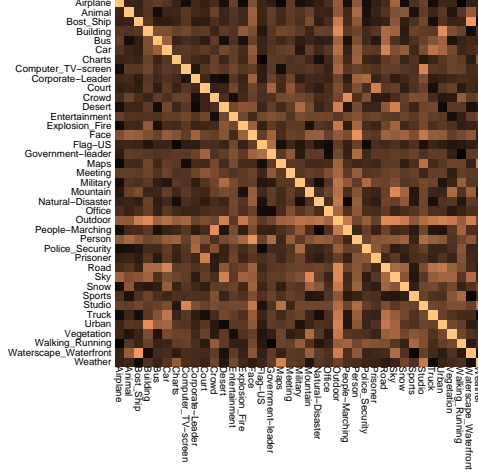


Fig. 4. The normalized mutual information between each pair of the 39 concepts in the LSCOM-Lite annotations data set. These are computed based on the annotations of the development data set in the experiments (see Section 5).

can be computed as

$$E(y_p | \mathbf{x}, \mathbf{y}_{\varphi \setminus p}^*) = P(y_p = +1 | \mathbf{x}, \mathbf{y}_{\varphi \setminus p}^*) - P(y_p = -1 | \mathbf{x}, \mathbf{y}_{\varphi \setminus p}^*)$$

where

$$P(y_p | \mathbf{x}, \mathbf{y}_{\varphi \setminus p}^*) = \frac{\exp\{-H(y_p \circ \mathbf{y}_{\varphi \setminus p}^* | \mathbf{x}, \mathbf{w})\}}{Z_p} = \frac{\exp\{F(\mathbf{x}, y_p \circ \mathbf{y}_{\varphi \setminus p}^*; \mathbf{w})\}}{Z_p} \quad (24)$$

and

$$Z_p(\mathbf{x}, \mathbf{y}_{\varphi \setminus p}^*) = \sum_{y_p \in \{+1, -1\}} \exp\{-H(y_p \circ \mathbf{y}_{\varphi \setminus p}^* | \mathbf{x}, \mathbf{w})\} \quad (25)$$

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

2.4 An Illustration: Interacting concepts

In Section 2.3, 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

$$\langle \theta(\mathbf{x}, \mathbf{y}), \theta(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}) \rangle = \langle \mathbf{x}, \tilde{\mathbf{x}} \rangle \sum_{1 \leq k \leq K} \delta \llbracket y_k = \tilde{y}_k \rrbracket + \sum_{(p,q) \in \mathcal{N}} \delta \llbracket y_p = \tilde{y}_p \rrbracket \delta \llbracket y_q = \tilde{y}_q \rrbracket \quad (26)$$

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

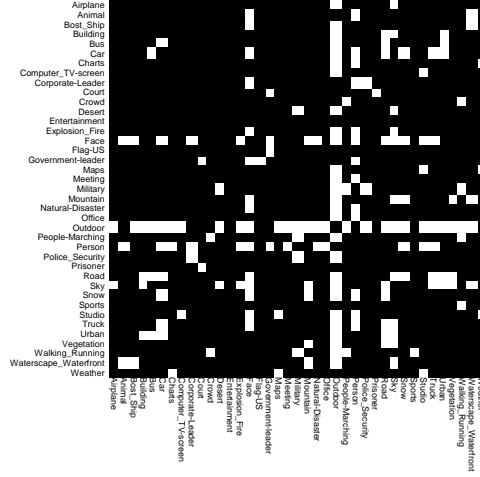


Fig. 5. The selected concept pairs according to the computed normalized mutual information. The white blocks indicate the selected concept pairs with significant correlations.

process in which the expensive expert labors are not required. First, we use the normalized mutual information [Yao 2003] to measure the correlations of each concept pair (p, q) as

$$NormMI(p, q) = \frac{MI(p, q)}{\min\{H(p), H(q)\}} \quad (27)$$

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

$$MI(p, q) = \sum_{y_p, y_q} P(y_p, y_q) \log \frac{P(y_p, y_q)}{P(y_p)P(y_q)} \quad (28)$$

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

$$H(p) = - \sum_{y_p \in \{+1, -1\}} P(y_p) \log P(y_p) \quad (29)$$

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 [Yao 2003], 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 4 illustrates the

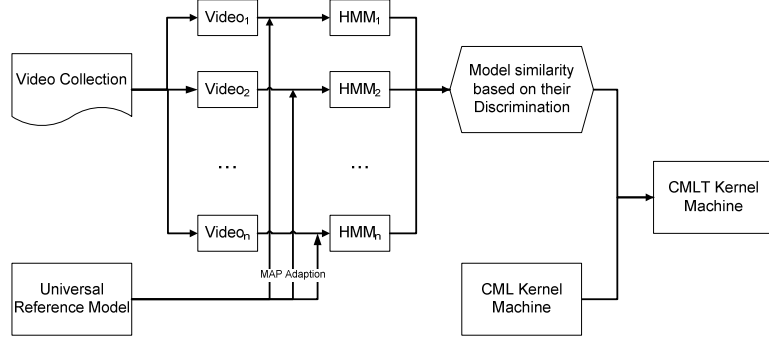


Fig. 6. The Correlative multi-label temporal kernel machine: It first adapts a universal reference model (URM) to a HMM for an individual video sequences. The model similarities can then be computed between these HMMs as temporal kernel based on their discrimination distances. By incorporating the temporal kernel into CML kernel machine, the CML temporal (CMLT) kernel machine can be obtained. Detailed algorithm is described in Section 3

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 5 represent the selected concept pairs.

3. CORRELATIVE MULTI-LABEL TEMPORAL KERNEL MACHINE FOR VIDEO ANNOTATION

In this section, we will introduce a temporal kernel machine under the above correlative multi-label video annotation framework. As aforementioned in Section 1.2, the temporal information of video sequences is an important source to characterize the inherent video dynamics when annotating video concepts, especially video event concepts. We will introduce a temporal kernel to represent the feature dynamic in video sequences.

In CML Kernel of Eqn. 5, we have indicated that the dot product $\langle \mathbf{x}, \tilde{\mathbf{x}} \rangle$ over the low-level feature vector \mathbf{x} and $\tilde{\mathbf{x}}$ can be substituted by any other kernel function. Therefore we can design a temporal-based kernel that characterizes the dynamics of video sequences. To design such a temporal kernel, our idea is first to design a distance measure $d(\mathbf{x}, \tilde{\mathbf{x}})$ between two videos $\mathbf{x}, \tilde{\mathbf{x}}$ and then a kernel can be computed through exponentiation as

$$K(\mathbf{x}, \tilde{\mathbf{x}}) = \exp \left\{ -\frac{d(\mathbf{x}, \tilde{\mathbf{x}})}{\sigma^2} \right\} \quad (30)$$

where σ is the kernel radius. As well known, the *Kullback-Leibler Divergence* (KLD) is a well-defined distance measure in information theory [Cover and Thomas 1991]. It can be used to compute the distribution distance between two statistical models. Therefore, if some dynamic models are constructed to capture the temporal dynamics of video sequences, KLD can then be computed between them. In this paper, we select Hidden Markov Models (HMMs) as such dynamic models. Specifically, for a video sequence (such as the subplot in this paper), we denote its observations as $O = \{o_t, t = 1, \dots, T\}$ where each o_t as the feature vectors for frame t in the video. Let there be Q states $\{1, \dots, Q\}$

and the state of each frame t is denoted by s_t . The transition probability $a_{i,j}$ denotes the state transition between the state i and j . For each state s_t , the observation o_t is generated according to an distribution $P(o_t|s_t)$. In this paper, we use *Gaussian Mixture Model* (GMM) as this observation distribution:

$$b_i(o_t) = P(o_t|s_t = i) = \sum_{l=1}^n \lambda_l^i \mathcal{N}(o_t|\mu_l^i, \Sigma_l^i) \quad (31)$$

where $\lambda_l^i, \mu_l^i, \Sigma_l^i$ is the mixing coefficient, the mean vector and covariance matrix of l th Gaussian component respectively, given the current state is i . For simplicity, the covariance matrix is assumed to be diagonal.

Given two video sequences and their respective HMMs $\Theta, \tilde{\Theta}$, we can compute the KLD [Cover and Thomas 1991] between them:

$$D_{KL}(\Theta||\tilde{\Theta}) = \int P(O|\Theta) \log \frac{P(O|\Theta)}{P(O|\tilde{\Theta})} \quad (32)$$

However, there exists no closed form expression for the KLD between these two HMMs. The most straightforward approach to computing this KLD is to use the Monte-Carlo simulation [Berg 2004]. But that will result in a significant computational cost. In this section, we will introduce an alternative approximation approach [Liu et al. 2007] that can be computationally more efficiently than the Monte-Carlo approach. It is aimed at computing an upper bound approximation of KLD between two HMMs. The approximation is motivated from the following upper bound that is based on the chain rule for relative entropy [Cover and Thomas 1991]:

LEMMA 1. *Given two mixture distributions $f = \sum_{i=1}^L w_i f_i$ and $g = \sum_{i=1}^L v_i g_i$, the KLD between them is upper bounded by*

$$D_{KL}(f||g) \leq D_{KL}(w||v) + \sum_{i=1}^L w_i D_{KL}(f_i||g_i) \quad (33)$$

where $D_{KL}(w||v) = \sum_{i=1}^L w_i \log \frac{w_i}{v_i}$. This inequality directly follows the log-sum inequality (see pp. 31 of [Cover and Thomas 1991]).

Let backward variables $\beta_t(i) = P(o_t o_{t+1} \cdots o_T | s_t = i, \Theta)$ denote the probability that the sequence $o_t o_{t+1} \cdots o_T$ is observed given the current state s_t is i and $\pi = [\pi_1 \pi_2 \cdots \pi_Q]^T$ denote the initial state distribution. Thus the distribution of the whole observation sequences can be computed by the Baum-Welch algorithm [Rabiner 1989] as

$$P(O|\Theta) = \sum_{i=1}^Q \pi_i \beta_t(i) \quad (34)$$

Therefore based on lemma, the KLD between two HMMs $\Theta, \tilde{\Theta}$ can be computed from Lemma as

$$\begin{aligned} D_{KL}(\Theta||\tilde{\Theta}) &= D_{KL}\left(\sum_{i=1}^Q \pi_i \cdot \beta_t(i) \middle| \middle| \sum_{i=1}^Q \tilde{\pi}_i \cdot \tilde{\beta}_t(i)\right) \\ &\leq D_{KL}(\pi||\tilde{\pi}) + \sum_{i=1}^Q \pi_i \cdot D_{KL}(\beta_t(i)||\tilde{\beta}_t(i)) \end{aligned} \quad (35)$$

The term $D_{KL}(\beta_t(i)||\tilde{\beta}_t(i))$ can be computed by utilizing the following recursive for-

mulation:

$$\beta_t(i) = b_i(o_t) \sum_{j=1}^Q a_{i,j} \beta_{t+1}(j) \quad (36)$$

Thus

$$D_{KL}(\beta_t(i) \|\tilde{\beta}_t(i)) \leq D_{KL}(b_i \|\tilde{b}_i) + D_{KL}(a_{i,\cdot} \|\tilde{a}_{i,\cdot}) + \sum_{i=1}^Q a_{i,j} D_{KL}(\beta_{t+1}(j) \|\tilde{\beta}_{t+1}(j)) \quad (37)$$

We can define $D_t = [D_t^1 D_t^2 \cdots D_t^Q]^T$ with $D_t^i = D_{KL}(\beta_t(i) \|\tilde{\beta}_t(i))$ and $C = [C_1 C_2 \cdots C_Q]^T$ with $C_i = D_{KL}(b_i \|\tilde{b}_i) + D_{KL}(a_{i,\cdot} \|\tilde{a}_{i,\cdot})$. Thus Eqn. (35) and (37) can then be rewritten as

$$\begin{aligned} D_{KL}(\Theta \|\tilde{\Theta}) &\leq D_{KL}(\pi \|\tilde{\pi}) + \pi^T D_1 \\ D_t &\leq C + A \cdot D_{t+1} \end{aligned} \quad (38)$$

where $A = (a_{i,j})_{Q \times Q}$ is the transition matrix. Therefore, we have

$$D_{KL}(\Theta \|\tilde{\Theta}) \leq D_{KL}(\pi \|\tilde{\pi}) + \pi^T \left(\sum_{t=1}^{T-1} A^{t-1} C + A^{T-1} D \right) \quad (39)$$

Assume that the model Θ is stationary so a stationary distribution γ exists, i.e.,

$$\begin{aligned} \gamma^T A &= \gamma^T \\ \lim_{t \rightarrow \infty} \pi^T A^t &= \gamma^T \end{aligned} \quad (40)$$

Therefore, combine the Eqn. (39) and (40), the KLD rate between two HMMs can be

$$\begin{aligned} \widehat{D}_{KL}(\Theta \|\tilde{\Theta}) &= \lim_{T \rightarrow \infty} \frac{1}{T} D_{KL}(\Theta \|\tilde{\Theta}) \\ &\leq \gamma^T C = \sum_{i=1}^Q \gamma_i \left\{ D_{KL}(b_i \|\tilde{b}_i) + D_{KL}(a_{i,\cdot} \|\tilde{a}_{i,\cdot}) \right\} \end{aligned} \quad (41)$$

Similarly, we can obtain the reverse KLD rate as

$$\widehat{D}_{KL}(\tilde{\Theta} \|\Theta) \leq \sum_{i=1}^Q \tilde{\gamma}_i \left\{ D_{KL}(\tilde{b}_i \|\tilde{b}_i) + D_{KL}(\tilde{a}_{i,\cdot} \|\tilde{a}_{i,\cdot}) \right\} \quad (42)$$

Where $\tilde{\gamma}$ is the stationary distribution of the model $\tilde{\Theta}$. So the symmetric KLD rate is

$$\begin{aligned} D(\Theta \|\tilde{\Theta}) &\leq \frac{1}{2} \sum_{i=1}^Q \gamma_i \left\{ D_{KL}(b_i \|\tilde{b}_i) + D_{KL}(a_{i,\cdot} \|\tilde{a}_{i,\cdot}) \right\} \\ &+ \frac{1}{2} \sum_{i=1}^Q \tilde{\gamma}_i \left\{ D_{KL}(\tilde{b}_i \|\tilde{b}_i) + D_{KL}(\tilde{a}_{i,\cdot} \|\tilde{a}_{i,\cdot}) \right\} \end{aligned} \quad (43)$$

Substitute the above upper bound of the symmetric KLD rate into Eqn. (30), we can obtain the temporal kernel between two video sequences as

$$\begin{aligned} K(\Theta, \tilde{\Theta}) &= \\ \exp \left\{ - \frac{\sum_{i=1}^Q \gamma_i \left\{ D_{KL}(b_i \|\tilde{b}_i) + D_{KL}(a_{i,\cdot} \|\tilde{a}_{i,\cdot}) \right\} + \sum_{i=1}^Q \tilde{\gamma}_i \left\{ D_{KL}(\tilde{b}_i \|\tilde{b}_i) + D_{KL}(\tilde{a}_{i,\cdot} \|\tilde{a}_{i,\cdot}) \right\}}{2\sigma^2} \right\} \end{aligned} \quad (44)$$

With the above temporal kernel, we can define the Correlative Multi-Label Temporal Kernel (CMLTK) by incorporating Eqn. (44) into Eqn. (5) as

$$K(\langle\theta(\mathbf{x}, \mathbf{y}), \theta(\tilde{\mathbf{x}}, \tilde{\mathbf{y}})\rangle) = \exp \left\{ - \frac{\sum_{i=1}^Q \gamma_i \{D_{KL}(b_i || \tilde{b}_i) + D_{KL}(a_{i,\cdot} || \tilde{a}_{i,\cdot})\} + \sum_{i=1}^Q \tilde{\gamma}_i \{D_{KL}(\tilde{b} || b_i) + D_{KL}(\tilde{a}_{i,\cdot} || a_{i,\cdot})\}}{2\sigma^2} \right\} \cdot \sum_{1 \leq k \leq K} \delta[y_k = \tilde{y}_k] + \sum_{1 \leq p < q \leq K} \delta[y_p = \tilde{y}_p] \delta[y_q = \tilde{y}_q] \quad (45)$$

Such a multi-label temporal kernel considers not only the concept interactions between each other but also the temporal evolution of video sequences. In this paper, we call Eqn. (45) by *Correlative Multi-Label Temporal (CMLT) Kernel*.

Finally, the KLD between the two GMMs distributions b_i, \tilde{b}_i in the above equations can be approximated through unscented transform [Goldberger and Aronowitz 2005].

$$D_{KL}(b_i || \tilde{b}_i) = \frac{1}{2d} \sum_{l=1}^n \lambda_l^i \sum_{k=1}^{2d} \log \frac{N(x_{l,k}^i | \mu_l^i, \Sigma_l^i)}{N(x_{l,k}^i | \tilde{\mu}_l^i, \tilde{\Sigma}_l^i)} \quad (46)$$

Where d is the dimension of the observed feature vectors, and $x_{l,k}^i$ is the ‘‘sigma’’ points defined as

$$\begin{aligned} x_{l,k}^i &= \mu_l^i + \left(\sqrt{d \Sigma_l^i} \right)_k, k = 1, \dots, d \\ x_{l,d+k}^i &= \mu_l^i - \left(\sqrt{d \Sigma_l^i} \right)_k, k = 1, \dots, d \end{aligned} \quad (47)$$

These sample points completely capture the true mean and variance of the Gaussian distribution $N(x | \mu_l^i, \Sigma_l^i)$, i.e., the l -th component distribution given its corresponding state is i .

3.1 A Universal Reference Model

As stated in Section 3.1, we use an upper bound to approximate the intractable exact KLD between two HMMs. These two models have the same state number Q . However, since they are trained independently on their own video sequences, the correspondence between their respective states may not be in the same order from 1 to Q . Such a disaccord between the states in the two models can lead to an upper bound that is not tight enough. To obtain a tighter bound, we can first train a *Universal Reference Model (URM)* from referential sequences, e.g., some video sequences from the training set. Then given a new video, its HMM can be adapted from this URM. Since the models are all adapted from this URM, the states will have a reasonable correspondence between the models. Thus, the obtained upper bound will be much tighter than that computed from the independently-trained models.

In this paper, the standard *Maximum A Posteriori (MAP)* technique [Gauvain and Lee 1994] is used to adapt the HMM. Formally, given the parameters Θ^{URM} of the URM and a new observation O of the new video sequence, we estimate the new HMM Θ . We use Θ^{URM} as the initial parameter. As suggested in [Gauvain and Lee 1994], the standard *Expectation-Maximization (EM)* algorithm is then applied to update Θ repeatedly until convergence except for the mean vector of GMMs, i.e.

$$\mu_l^i \leftarrow \alpha \cdot \mu_l^i + (1 - \alpha) \cdot \frac{\sum_{t=1}^T o_t \cdot P(s_t = i, m_t^i = l | O, \Theta)}{\sum_{t=1}^T P(s_t = i, m_t^i = l | O, \Theta)} \quad (48)$$

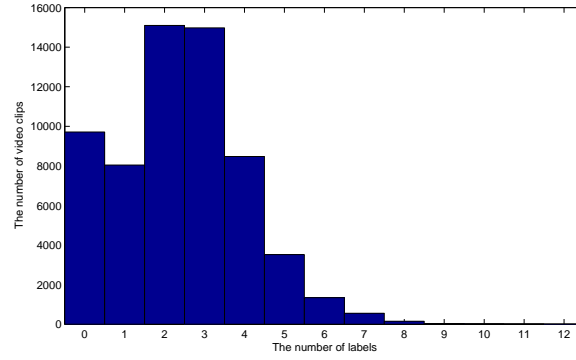


Fig. 7. The numbers of labels for the video clips in LSCOM-Lite Annotation data set.

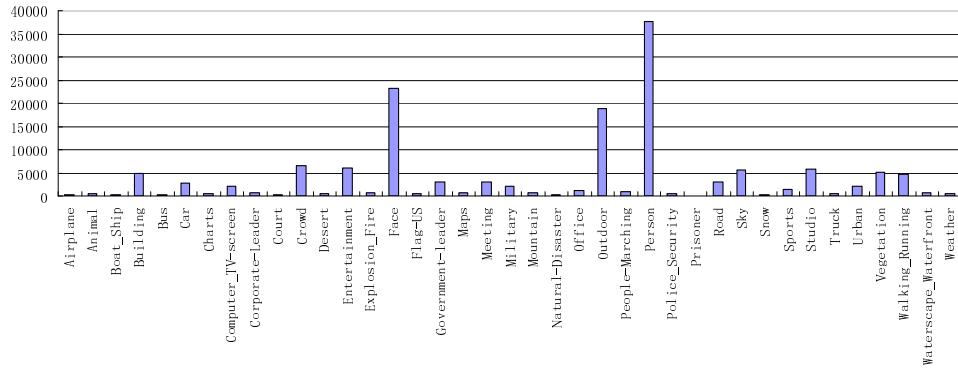


Fig. 8. Video Concepts and their distribution in LSCOM-Lite data set

where m_t^i indicates the mixture component given the state is i at time slice t and α is the weighting factor giving the bias between the previous estimate and the current one. We will set α to be 0.7 in the experiment. The update rules for all the other parameters follow the EM algorithm.

4. EXPERIMENTS

In this section, we conduct the proposed algorithms on the widely used benchmark TRECVID dataset. We will show the experimental results on two proposed kernel machines. (1) the multi-label kernel machine described in section 2. It exploits the individual concepts and their correlations in a single CML kernel. (2) the multi-label temporal kernel machine described in section 3. It incorporates the temporal information into CML kernel and models the concept interactions and low-level feature dynamics in CMLT kernel together. We will compare them with other state-of-the-art methods in the first and second paradigms.

4.1 TRECVID Set Description

To evaluate the proposed video annotation algorithm, we conduct the experiments on the benchmark TRECVID 2005 data set [TRECVID]. This is one of the most widely used data sets by many groups in the area of multimedia concept modeling [Campbell and et al. 2006][Chang and et al. 2006][Hauptmann and et al. 2006]. 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.

For each subshot, 39 concepts are multi-labeled according to LSCOM-Lite annotations [Naphade et al. 2005]. These annotated concepts consist of a wide range of genres, including program category, setting/scene/site, people, object, activity, event, and graphics. Figure 8 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 6).

Figure 7 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.

4.2 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 2.) 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 [Godbole and Sarawagi 2004]. 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 $\langle \mathbf{x}, \tilde{\mathbf{x}} \rangle$ 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

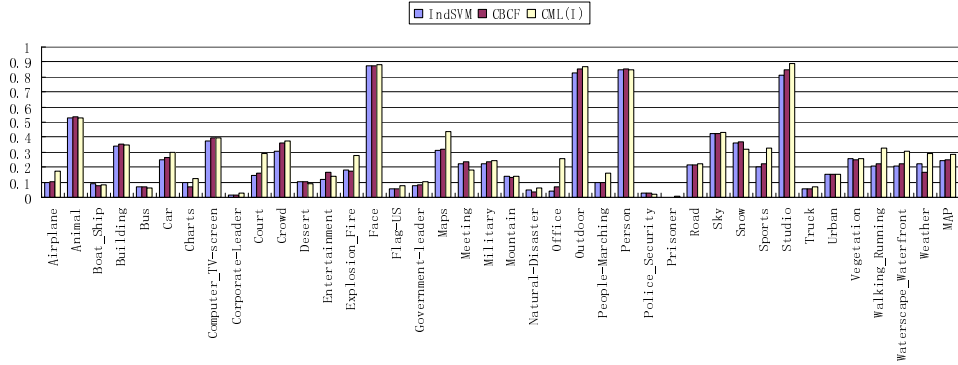


Fig. 9. The performance comparison of IndSVM, CBCF and CML(I).

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.

4.3 Experiment One: Correlative Multi-Label Kernel Machine

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. (26) is used, and this approach is denoted by CML(II).

All subshots are then processed to extract several kinds of low-level features on the keyframes of these subshots [Hua et al. 2006], including

1. Block-wise Color Moment in Lab color space (225D): based on 5-by-5 division of images in Lab color space;
2. Co-occurrence Texture (20D);
3. Wavelet Texture (128D);
4. Edge Distribution Layout (75D);

and some mid-level features

5. Face (7D): consisting of the face number, face area ratio, the position of the largest face.

Since these features are extracted statically on only keyframes, they are called by Static Features (SF), which are different from the Dynamic Features (DF) used in temporal kernel (Eqn. (44)).

4.3.1 Experiment I: Fully-Correlative Concepts. We first conduct experiments of the multi-label method CML (I) with the fully-correlative concepts. It considers all possible correlations between the concepts. Figure 9 illustrates the performance of CML(I) compared to that of IndSVM (first paradigm) and CBCF (second paradigm). The following observations can be obtained:

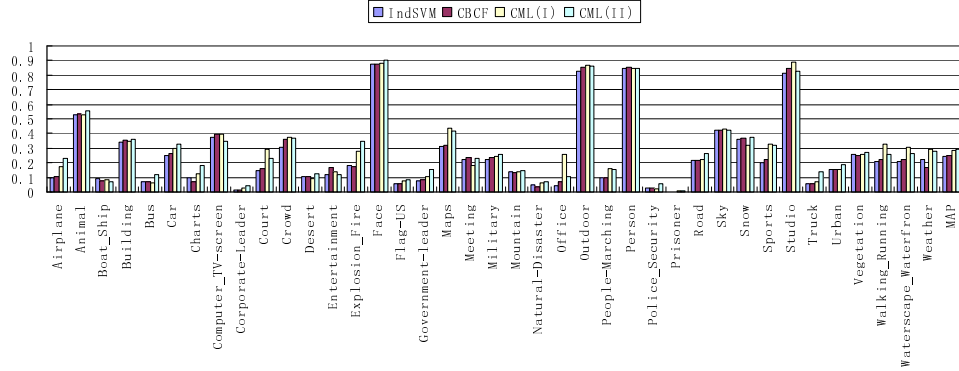


Fig. 10. The performance comparison of IndSVM, CBCF, CML (I) and CML(II).

- 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 IndSVM 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.

4.3.2 *Experiment II: Partially-Correlative Concepts.* Following the proposed approach in Section 2.4, the deterioration problem can be solved by removing concept pairs with insignificant correlations.

Figure 6 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 5 shows these selected concept pairs. As we can see, these preserved concept pairs either have intuitive semantic correlations e.g. “waterscape water-front” 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 10 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.

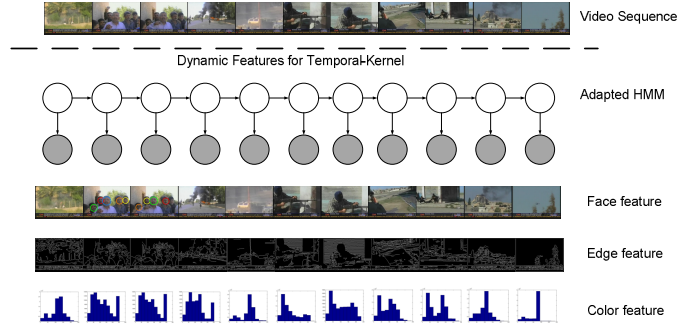


Fig. 11. The dynamic features used in temporal kernel: the low-level features are extracted at the rate of one frame per second, and these extracted features are then used to train an adapted HMM from URM. It is contrast to the static features that are extracted only on keyframes of the subshots.

—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.

4.4 Experiment Two: Multi-Label Temporal Kernel Machine

In this section, we evaluate the proposed CMLT kernel method in Section 3. As aforementioned, this method can further capture the temporal information of video sequences. Compared to the formal CML method that only extracts static features on the keyframes of video subshots, this CMLT kernel machine can capture the dynamic features contained in the temporal patterns of the videos. Such dynamic patterns are important sources for improving the discrimination between different video concepts.

As depicted in Section 3, all subshots are seen as a sequence of video frames, and the low-level features are extracted on these frame sequences rather than only keyframes of each subshot. To accelerate the feature extraction and model learning, we do not extract features on every frame. Instead, we only extract the features at the rate of one frame per second. These extracted features are then used to train the HMM for each subshot. In more detail, A universal reference model is first trained on 5000 video subshots which are randomly selected from the training set. Then for each subshot, a HMM is adapted from this URM according to Eqn. (48) and EM algorithm (see Section 3.2 for detail). The low-level features extracted on the video frames are the same as the static features used in experiment one. However, since they are extracted on frame sequences to train a dynamic model, we call them *Dynamic Features* (DF) (see Figure 11).

For the sake of the fair comparison, we follow the same experiment settings in experiment one. Table 1 illustrates the performance of CMLT kernel method with the comparisons of IndSVM, CBCF, CML(I), CML(II). From these results, we can find

—The CMLT machine has the best overall performance in terms of MAP. It outperforms the IndSVM, CBCF, CML(I) and CML(II) by 35.0%, 31.3%, 17.0%, 14.7%.

	IndSVM	CBCF	CML(I)	CML(II)	CMLT
Airplane	0.1005	0.1019	0.1712	0.2325	0.2563
Animal	0.5265	0.5336	0.5302	0.55824	0.7193
Boat_Ship	0.087	0.0798	0.0849	0.0707	0.0779
Building	0.3375	0.3538	0.3486	0.3585	0.3952
Bus	0.0669	0.0724	0.0602	0.1147	0.1706
Car	0.2469	0.2673	0.2983	0.3296	0.4185
Charts	0.0981	0.0709	0.1277	0.182	0.2558
Computer_TV-screen	0.3773	0.3927	0.3976	0.3438	0.379
Corporate-Leader	0.0112	0.014	0.03	0.0438	0.0483
Court	0.1462	0.1568	0.294	0.232	0.2558
Crowd	0.3073	0.3598	0.3775	0.3676	0.4053
Desert	0.1047	0.1053	0.0902	0.125	0.1378
Entertainment	0.1174	0.1687	0.14	0.1171	0.1291
Explosion_Fire	0.1773	0.1768	0.2755	0.3447	0.38
Face	0.8779	0.8782	0.8854	0.9062	0.9762
Flag-US	0.0571	0.0563	0.0759	0.084	0.0926
Government-leader	0.0774	0.0838	0.1029	0.1515	0.167
Maps	0.3147	0.3206	0.4347	0.4156	0.5228
Meeting	0.2208	0.2391	0.183	0.232	0.2558
Military	0.2202	0.2337	0.2405	0.2571	0.2394
Mountain	0.1367	0.135	0.1397	0.148	0.1632
Natural-Disaster	0.0462	0.0381	0.0633	0.0664	0.0932
Office	0.044	0.0706	0.2541	0.1053	0.1161
Outdoor	0.823	0.8517	0.8695	0.8607	0.8166
People-Marching	0.095	0.0998	0.1595	0.1561	0.2949
Person	0.8441	0.8535	0.8453	0.844	0.9856
Police_Security	0.0301	0.0253	0.02	0.058	0.0639
Prisoner	0.0026	0.0016	0.0039	0.0096	0.0106
Road	0.2169	0.2158	0.2249	0.2656	0.2928
Sky	0.4204	0.4261	0.4281	0.4213	0.4645
Snow	0.3625	0.37	0.3179	0.374	0.4123
Sports	0.2025	0.2194	0.329	0.3226	0.3998
Studio	0.8109	0.8448	0.889	0.8283	0.9132
Truck	0.0552	0.0529	0.0727	0.1381	0.1523
Urban	0.151	0.1528	0.1517	0.1861	0.2052
Vegetation	0.2596	0.2511	0.2537	0.2675	0.2949
Walking_Running	0.2094	0.2188	0.3251	0.2565	0.3828
Waterscape_Waterfront	0.2049	0.2219	0.3055	0.2642	0.2913
Weather	0.22	0.169	0.2898	0.2765	0.3364
MAP	0.2463	0.2534	0.2843	0.2901	0.3326

Table I. The average precision over 39 LSCOM-lite concepts for the five algorithms: IndSVM, CBCF, CML(I), CML(II), CMLT. The CMLT gains the best over performance of these algorithm, and it also outperforms the other four algorithms on 30 out of 39 concepts.

—CMLT gains the best performance on 30 concepts out of the whole 39 concepts.

—Moreover, on four event-related concepts, i.e., “Explosion_Fire”, “Natural-Disaster”, “People-Marching”, “Walking_Running”, the CMLT outperforms the other four methods, because it takes advantage of the temporal dynamics contained in these event concepts.

5. CONCLUSION

We propose a *Correlative Multi-Label* (CML) kernel machine in this paper to leverage the label correlations to help infer the video concepts. It exploits the individual concepts and their correlations in a single formulation. Furthermore, a temporal kernel is proposed to be incorporated into the CML formulation to form a *Correlative Multi-Label Temporal* (CMLT) kernel machine. This new kernel method takes into account not only feature dynamics but also concept interactions in an integrated manner. It obeys the *principle of least commitment* without any extra step that can propagate errors to its consecutive step. Experiment on benchmark TRECVID data set demonstrates the significant improvement is obtained compared to the state-of-the-art algorithms in the other two paradigms for video annotation.

REFERENCES

- BERG, B. A. 2004. *Markov chain Monte Carlo simulations and their statistical analysis*. World Scientific.
- BOYD, S. AND VANDENBERGHE, L. 2004. *Convex Optimization*. Cambridge University Press.
- CAMPBELL, M. AND ET AL. 2006. Ibm research trecvid-2006 video retrieval system. In *TREC Video Retrieval Evaluation (TRECVID) Proceedings*.
- CHANG, S.-F. AND ET AL. 2006. Columbia university trecvid-2006 video search and high-level feature extraction. In *TREC Video Retrieval Evaluation (TRECVID) Proceedings*.
- COVER, T. AND THOMAS, J. 1991. *Elements of information theory*. John Wiley and Sons, New York, USA.
- CRISTIANINI, N. AND SHAWE-TAYLOR, J. 2000. *An introduction to support vector machines and other kernel-based learning methods*. Cambridge University.
- EBADOLLAHI, S., XIE, L., CHANG, S.-F., AND SMITH, J. R. 2006. Visual event detection using multi-dimensional concept dynamics. In *IEEE International Conference on ICME*.
- GAUVAIN, J.-L. AND LEE, C.-H. 1994. Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains. *IEEE Transaction on Speech and Audio Processing* 2, 2, 291–298.
- GODBOLE, S. AND SARAWAGI, S. 2004. Discriminative methods for multi-labeled classification. In *PAKDD*.
- GOLDBERGER, J. AND ARONOWITZ, H. 2005. A distance measure between gmms based on the unscented transform and its application to speaker recognition. In *INTERSPEECH*.
- HAUPTMANN, A., CHEN, M.-Y., AND CHRISTEL, M. 2004. Confounded expectations: Informedia at TRECVID 2004. In *TREC Video Retrieval Evaluation Online Proceedings*.
- HAUPTMANN, A. G. AND ET AL. 2006. Multi-lingual broadcast news retrieval. In *TREC Video Retrieval Evaluation (TRECVID) Proceedings*.
- HUA, X.-S., MEI, T., LAI, W., WANG, M., TANG, J., QI, G.-J., LI, L., AND GU, Z. 2006. Microsoft reasech asia trecvid 2006 high-level feature extraction and rushes exploitation. In *Online proc. of the TRECVID workshops*.
- JIANG, W., CHANG, S.-F., AND LOUI, A. 2006. Active concept-based concept fusion with partial user labels. In *Proceedings of IEEE International Conference on Image Processing*.
- LAFFERTY, J., MCCALLUM, A., AND PEREIRA, F. 2001. Conditional random fields: Probabilistic models for segmentation and labeling sequence data. In *Proc. of International Conference on ICML*.
- LIU, P., SOONG, F. K., AND ZHOU, J.-L. 2007. Divergence-based similarity measure for spoken document retrieval. In *IEEE International Conference on ICASSP*.
- MARR, D. 1982. *Vision*. W.H.Freeman and Company.
- NAPHADE, M., KOZINTSEV, I., AND HUANG, T. 2002. Factor graph framework for semantic video indexing. *IEEE Trans. on CSVT* 12, 1 (Jan.).
- NAPHADE, M. R. 2002. Statistical techniques in video data management. In *IEEE Workshop on Multimedia Signal Processing*.
- NAPHADE, M. R., KENNEDY, L., KENDER, J. R., CHANG, S.-F., SMITH, J. R., OVER, P., AND HAUPTMANN, A. 2005. A light scale concept ontology for multimedia understanding for TRECVID 2005. In *IBM Research Report RC23612 (W0505-104)*.
- NIGAM, K., LAFFERTY, J., AND MCCALLUM, A. 1999. Using maximum entropy for text classification. In *IJCAI-99 Workshop on Machine Learning for Information Filtering*. 61–67.

- RABINER, L. R. 1989. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE* 77, 2, 257–286.
- SMITH, J. R. AND NAPHADE, M. 2003. Multimedia semantic indexing using model vectors. In *Proceeding of IEEE International Conferences on Multimedia and Expo*.
- SNOEK, C. G. M., WORRING, M., GEMERT, J. C., GEUSEBROEK, J.-M., AND SMEULDERS, A. W. M. 2006. The challenge problem for automated detection of 101 semantic concepts in multimedia. In *Proceedings of the ACM International Conference on Multimedia*. Santa Barbara, USA, 421–430.
- TANG, J., HUA, X.-S., QI, G.-J., WANG, M., MEI, T., AND WU, X. 2007. Structure-sensitive manifold ranking for video concept detection. In *ACM International Conference on Multimedia*.
- TRECVID. <http://www-nlpir.nist.gov/projects/trecvid/>.
- TSOCHANTARIDIS, I., HOFMANN, T., JOACHIMS, T., AND ALTUN, Y. 2004. Support vector machine learning for interdependent and structured output spaces. In *Proc. of International Conference on ICML*.
- WANG, T., LI, J., DIAO, Q., HU, W., ZHANG, Y., AND DULONG, C. 2006. Semantic event detection using conditional random fields. In *Proc. of IEEE CVPR Workshop*.
- WINKLER, G. 1995. *Image analysis, random fields and dynamic Monte Carlo methods: A mathematical introduction*. Springer-Verlag, Berlin, Heidelberg.
- WU, Y., TSENG, B. L., AND SMITH, J. R. 2004. Ontology-based multi-classification learning for video concept detection. In *Proceeding of IEEE International Conferences on Multimedia and Expo*.
- XIE, L. AND CHANG, S.-F. 2002. Structural analysis of soccer video with hidden markov models. In *IEEE International Conference on ICASSP*.
- YAO, Y. Y. 2003. *Entropy measures, maximum entropy principle, and emerging applications*. Springer, Chapter Information-theoretic measures for knowledge discovery and data mining, 115–136.
- ZHA, Z.-J., MEI, T., HUA, X.-S., QI, G.-J., AND WANG, Z. 2007. Refining video annotation by exploiting pairwise concurrent relation. In *ACM International Conference on Multimedia*.

Received January 2008; accepted XX XXXX