Bi-Level Semantic Representation Analysis for Multimedia Event Detection

Xiaojun Chang, Zhigang Ma, Yi Yang, Zhiqiang Zeng, and Alexander G. Hauptmann

Abstract—Multimedia event detection has been one of the major endeavors in video event analysis. A variety of approaches have been proposed recently to tackle this problem. Among others, using semantic representation has been accredited for its promising performance and desirable ability for human-understandable reasoning. To generate semantic representation, we usually utilize several external image/video archives and apply the concept detectors trained on them to the event videos. Due to the intrinsic difference of these archives, the resulted representation is presumably to have different predicting capabilities for a certain event. Notwithstanding, not much work is available for assessing the efficacy of semantic representation from the source-level. On the other hand, it is plausible to perceive that some concepts are noisy for detecting a specific event. Motivated by these two shortcomings, we propose a bi-level semantic representation analyzing method. Regarding source-level, our method learns weights of semantic representation attained from different multimedia archives. Meanwhile, it restrains the negative influence of noisy or irrelevant concepts in the overall concept-level. In addition, we particularly focus on efficient multimedia event detection with few positive examples, which is highly appreciated in the real-world scenario. We perform extensive experiments on the challenging TRECVID MED 2013 and 2014 datasets with encouraging results that validate the efficacy of our proposed approach.

Index Terms—Multimedia Event Detection (MED), Semantic Representation, Bi-Level, Source-Level, Concept-Level.

1 INTRODUCTION

Multimedia content analysis has been progressing from simple concepts [8][6] to complicated events [7]. The growing research effort is of great interest to improve user experience in multimedia search, particularly video search. As promoted by the TRECVID, multimedia event detection (MED) has been one of the major endeavors in video event analysis. Per definition, MED is to detect the occurrence of an event within a video clip based on an Event Kit, which contains some text description about the concept and some example videos. Noteworthily, all the events in MED task are more complicated than those focused in news, sports or surveillance video analysis.

Existing MED systems aim to gain good detection performance with different number of positive examples [29], [31]. Shirahama et al. have presented a Hidden Conditional Random Field based algorithm using 100 positive training examples for MED [40]. Ma et al. have proposed to adapt knowledge from auxiliary domain for MED when only 10 positive training exemplars are provided [34]. Jiang et al. have proposed a self-paced re-ranking algorithm for zero-shot MED [25]. Among them, using 10 positive examples in the training data is particularly interesting since this scenario well matches what we expect from a practical system. In such a system, a user selects a few positive examples and negative examples, trains a classifier and then applies it to any unseen videos to get event detection result in a timely manner. Hence, in this paper we focus on effective MED using 10 positive training examples with high efficiency, achieved by a linear approach.

On the other hand, an MED system is desired to be intelligent with human-understandable reasoning that provides a list of concepts relevant to the target event. With such motivation, a few researchers have worked on using semantic representation for MED. These findings demonstrate that semantic representation is comparable to hand-crafted low-level features in performance while has the advantage of identifying event-related concepts for reasoning. Technically, semantic representation for event videos is obtained by training a variety of concept detectors related to objects, scenes and actions, etc. and then applying these detectors to the videos [23]. For current MED task, the concept detectors are usually trained using multiple image/video archives [51]. These archives may have disparate structures leading to semantic representations with different capabilities for event detection. However, limited work has investigated how to more effectively exploit such
difference within the semantic representations for improved MED. Additionally, it is reasonable to presume that some concepts within the semantic representations are irrelevant and noisy to a target, making them useless for prediction [8]. For example, a concept *swimming* is unlikely to be useful for detecting the event *Tuning a musical instrument*. Uncovering such irrelevance is also overlooked in the literature. Therefore, we propose a bi-level model for MED to capture the different predicting abilities derived from the source-level while restraining the effect of irrelevant concepts in the overall concept-level.

Specifically, we consider the representations learned from multiple image/video sources both as multiple groups and as an overall semantic representation simultaneously. The first task is to characterize the different predicting capabilities of different groups. Assuming that the capabilities can be characterized by how well the concept-based representation is structured, we propose to explore the interaction between different concepts within each group. Inspired by recent attention on feature interaction [44][19] in mining the data structure, we apply it to our problem and calculate the semantic interaction within each group as a way to differentiate the predicting abilities. Meanwhile, a sparse model [18] is utilized to uncover the useless concepts within the overall semantic representation combined from all the groups. The two schemes are integrated into a unified framework that is able to learn weights for different semantic representation groups and suppress the influence of noisy concepts.

Figure 1 illustrates the framework of our method. Multiple image/video sources are used to train the concept detectors that are subsequently applied to the MED videos to generate multiple groups of semantic representation. Next, the semantic interaction within each group is explored to learn the weights accordingly. At the same time, we leverage sparse learning on the overall combined semantic representation. The two processes are embedded into the detector learning, for which we have the predictions from multiple groups of semantic representation as well as the prediction from the overall semantic representation. The predictions are supposed to be consistent. Simultaneously, the prediction from the overall representation is constrained to fit the uncovered structure based on semantic interaction. These different components are cast into a joint framework so that they are tightly coupled. The optimized classification models from different groups of semantic representation together with the weights are used for prediction on testing videos. We name our method Bi-Level Semantic Analysis (B-LSA).

The main contributions of our work are three-fold:
- We propose to investigate the varying contribution of semantic representations from different image/video sources, thus enhancing the exploitation of semantic representation in the source-level.
- The noisy/irrelevant concepts within the overall semantic representation is restrained during the learning of our framework. In this way, the classification model is optimized in the concept-level.
- The source-level and concept-level exploration is integrated to make the two modules mutually beneficial and reciprocal.

2 RELATED WORK

In this section, we give a general review of multimedia event detection, then particularly introduce the progress on MED with few positive examples and on MED using semantic representation.

In 2010, TRECVID launched the MED challenge aiming to encourage new technologies for detecting more complicated multimedia events, e.g., *making a sandwich*. Ever since, researchers have been making effort to improve MED from different perspectives. Jiang et al. have proposed to use high-level features atop low-level ones by encoding high-level features into graphs and diffusing the scores obtained from low-level
features on the established graph for MED [24]. Ye et al. [47] have presented a new video representation built upon joint audio-visual bi-modal words. The proposed bi-modal words are obtained by exploring the relation across the quantized words extracted from both the visual and audio modalities in video. A few pooling strategies are subsequently harnessed to formulate the bi-modal into the final video representation. Ma et al. [36] have proposed to treat the negative videos unequally by assigning fine-grained labels to them in their classification model, motivated by that many negative videos may resemble the positive videos in different degrees. A joint model is developed to optimize the fine-grained labels with the video representations for a more robust classifier. Younessian et al. have utilized multi-modal knowledge consisting of Automatic Speech Recognition (ASR) transcripts, acoustic concept indexing and visual semantic indexing for MED [49]. Liu et al. have studied how to use the high-level face information to assist in MED considering that people are usually the central subjects in videos [32]. Yu et al. have proposed a way to learn abstract semantic "regions" during the feature pooling, thus being able to remove the "junk" information in video for better detection performance [50]. Ma et al. have proposed an approach that exploits the external concept-based videos and target event videos simultaneously to learn an intermediate representation coupled with the event classifier [35]. Gkalelis et al. have combined a novel nonlinear discriminant analysis method and a linear Support Vector Machine to improve the computational efficiency for MED [21].

Given that precisely labeled positive examples are difficult to get in a real system, a few researchers have investigated how to gain reasonable performance with few positive examples. A common strategy is to utilize auxiliary resources to compensate the scarce information from the few positive examples. Ma et al. have introduced a multi-task learning model to adapt knowledge from external image/video databases to the target event videos for learning a better event detector with only 10 positive training examples [34]. Habibian et al. have leveraged textual descriptions of web videos to learn an embedding as the representation of MED videos in few-example setting [22]. Some researchers have also advocated using related videos to improve MED with 10 positive training examples. Related videos are those closely but not fully associated with the event of interest. Yang et al. have presented an algorithm that automatically evaluates how positive the related exemplars are for the detection of an event and uses them on an exemplar-specific basis for learning the event classification model [46]. A weighted margin SVM formulation is developed by Tzelepis et al. to effectively incorporate related class observations for learning the classifier [45]. Notwithstanding, much effort is still needed in this realm since it is more consistent with the practical requirement regarding labeled training data.

On the other hand, some work has been focused on enabling the MED system to provide users with evidences that justify the prediction of a certain event. Usually, it is achieved by using semantic representation of event videos that is associated with concepts of objects, scenes and actions. Merler et al. have created a set of semantic classifiers trained from thousands of labeled web images, based on which they generate the semantic model vectors representation for MED [38]. Althoff et al. [9] have proposed a novel representation named Detection Bank learnt from a large number of object detectors. The Detection Bank is combined with a few state-of-the-art descriptors with prominent enhancement for MED. Mazloom et al. [37] have devised a solution using cross-entropy optimization to find the optimal concepts useful for a target event from a large concept pool. Can et al. have worked on mining the dependency of object-based concepts in the video level using a Markov Random Field based model, leading to a discriminative representation for MED [17]. A comprehensive research by Habibian et al. has explored which concepts should be adopted for the semantic representation, specifically targeting the number, the type, the specificity and the quality of the concept detectors [23].

Though existing work on semantic representation for MED has varying degree of success, not much research has been established to assess the efficacy of the representation in the source-level. Additionally, it is beneficial to restrain the negative influence of noisy or irrelevant concepts w.r.t. a certain event when predicting in the overall concept-level. Hence, we focus on addressing these two issues by developing an integrated framework. Further, we apply our method on MED with 10 positive examples considering their paucity in a real system.

## 3 Problem Formulation

This section contains the detailed formulation of our algorithm for better usage of semantic representation from multiple sources for MED. To make it clear, we base the illustration on Figure 1.

Suppose we have $m$ multimedia sources, indicated by "source 1" to "source $m$" in Figure 1. These sources are used to learn the semantic representation of the event videos, which is indicated by the orange arrow of Figure 1 and the event videos are indicated by "MED Videos". Learning the semantic representation is a process of training concept detectors and then applying them to the event videos to get the detection scores, which serve as the semantic representation. The representation of $n$ training videos is therefore denoted as $X_{s|m} = \left[ X_s; X_{s+1}; \ldots; X_m \right] \in \mathbb{R}^{d \times n}$, $d_s$ is the number of the semantic concepts of the $s$-th source. The concatenated representation is denoted as $X = [X_1; X_2; \ldots; X_m] \in \mathbb{R}^{d \times n}$ where $d = \sum_{s=1}^{m} d_s$. Let $x_i$ indicate one datum and $y \in \{0, 1\}^{n \times 1}$ be the label vector. $y_i = 1$ if $x_i$ is a positive example of the event whereas $y_i = 0$ otherwise.

We start formulating our problem from a general process of classifier training since event detection is essentially a classification problem. The classical regularized empirical error is used to learn a classifier $f$ based on a set of training data $\{x_i, y_i\}_{i=1}^{n}$ where $y_i$ indicates the label of $x_i$:

$$\min_{f} \sum_{i=1}^{n} \text{loss}(f(x_i), y_i) + \beta \Omega(f).$$

(1)

$\text{loss}(\cdot, \cdot)$ is a loss function and $\Omega(f)$ is the regularization
function on $f$ with $\beta$ as its parameter. We define $f(x_i)$ as:

$$f(x_i) = w^T x_i + b_i$$

which is often used in linear regression loss [10][11][12]. Here $w$ is a projection vector to be used for classification and $b_i$ is a bias term.

Adapting to the data matrices $X$ and $X_s^{(m)}_{i=1}$ defined above, we can easily integrate Eq. (1) and Eq. (2) into:

$$\min_{w,b,f,w_s,b_s,f} \left\| X^T w + 1_n b^T - f \right\|_2^2$$

$$+ \sum_{s=1}^m \left\| X_s^T w_s + 1_n b_s^T - f_s \right\|_2^2 + \beta (\Omega(w) + \Omega(w_s))$$

where $1_n \in \mathbb{R}^n$ denotes a column vector with all ones; $f$ and $f_s$ are introduced as the predicted label vectors that are useful for our proposed bi-level learning. $\Omega(w)$ and $\Omega(w_s)$ are regularization on $w$ and $w_s$ respectively for mining the sparsity of $w$ and avoiding over-fitting of $w_s$, which is elaborated as follows.

From the overall level, some concepts may not useful for characterizing a target event. For example, concepts like “fish” follow. One condition does not apply. Thus, we arrive at:

$$\min_{w,b,f,w_s,b_s,f} \left\| X^T w + 1_n b^T - f \right\|_2^2$$

$$+ \sum_{s=1}^m \left\| X_s^T w_s + 1_n b_s^T - f_s \right\|_2^2 + \lambda (\Omega(w) + \Omega(w_s))$$

where $\lambda \in \mathbb{R}$ is a parameter.

Next, we propose to regulate the predicted label $f$ to be smooth on the overall semantic interaction and consistent with $f_s$ because we want the learning models from each part to agree with each other as much as possible considering they are intrinsically predicting the same event. This process is illustrated by the modules “Prediction Fitting” and “Prediction Consistency” in our framework in Figure 1. After imposing this constraint, we arrive at the following objective function:

$$\min_{w,b,f,w_s,b_s,f} \left\| X^T w + 1_n b^T - f \right\|_2^2$$

$$+ \sum_{s=1}^m \left\| X_s^T w_s + 1_n b_s^T - f_s \right\|_2^2 + \lambda (\Omega(w) + \Omega(w_s))$$

As discussed by the aforementioned process, our objective function Eq. (5) is an integration of semantic interaction from each group of semantic representation, sparsity mining over the whole semantic representation, prediction consistency between different semantic groups, prediction fitting on the overall semantic representation and weight learning regarding different semantic groups.

4 Optimization

Through the optimization, we expect to get the optimal value of the event detectors $w_i$ from different groups of semantic representation and the corresponding weights $\lambda_s$. The process also includes the optimization of other variants $f_i$, $b_i$, $w$, $f$ and $b$. Notice that the “sparsity mining” component of our objective function involves $\ell_2,1$-norm which is non-smooth. Hence, we propose an alternating solution to solve it.

First, we fix the weights $\lambda_s$ and optimize $w_i$, $f_i$, $b_i$, $w$, $f$ and $b$. Notice that the “sparsity mining” component of our objective function involves $\ell_2,1$-norm which is non-smooth. Hence, we propose an alternating solution to solve it.

By setting the derivative w.r.t. $b_s$ to 0, we have:

$$b_s^* = \frac{1}{n} (1_n^T f_s - 1_n^T X_s^T w_s)$$

where $1_n \in \mathbb{R}^{n \times 1}$ is a column vector with all ones.

Let $H = I - \frac{1}{n} 1_n 1_n^T$ be a centering matrix where $I \in \mathbb{R}^{n \times n}$ is an identity matrix. The problem of Eq. (5) equals to the
Algorithm 1: Optimization procedure for B-LSA.

Input:

\[ X_s \in \mathbb{R}^{d \times n}, \quad X \in \mathbb{R}^{d \times n} \text{ and } y \in \mathbb{R}^{n \times 1}; \]
\[ \tilde{R}_s \in \mathbb{R}^{n \times n}; \]
Parameters \( \alpha, \beta \) and \( p \).

Output:

Optimized \( w_s \in \mathbb{R}^{d \times 1} \) and \( y \in \mathbb{R}^{n \times 1} \), \( b \in \mathbb{R}^{n \times 1} \) and \( \alpha \).

1: set \( t = 0 \), initialize \( \lambda^0 = \frac{1}{n} \) and \( w \in \mathbb{R}^{d \times 1} \),

Compute \( K_s = (X_s H X_s^T + \lambda I_s)^{-1} \);

Compute

\[ J_s = ((X_s^T K_s X_s - I_n)^2 + \alpha I_n + \beta H X_s^T K_s^2 X_s H)^{-1}; \]

repeat

Compute the diagonal matrix \( D^{t+1} \) according to \( D^{t+1} = \frac{1}{2||w||^2}; \)

Compute \( M^{t+1} = (X H X^T + \beta D^{t+1})^{-1}; \)

Update \( f^{t+1} \) according to Eq. (15);

Update \( w^{t+1} \) according to Eq. (16);

Update \( b^{t+1} \) according to Eq. (19);

Update \( t = t + 1 \);

until Convergence: \( |obj_{t+1} - obj_t| / obj_t \leq 10^{-3} \);

3: Return \( w_s, b_s, w, b \) and \( \lambda_s \).

following one when optimizing the event detectors \( w_s \):

\[ \min_w \sum_{s=1}^{m} \text{Tr} \left( (HX_s^T w_s - H f_s)^T (HX_s^T w_s - H f_s) \right) \]
\[ + \beta \sum_{s=1}^{m} \text{Tr} (w_s^T w_s) \]
\[ \Rightarrow \min_w \sum_{s=1}^{m} \text{Tr} (w_s^T (X_s H X_s^T + \beta I_s) w_s - 2w_s^T X_s H f_s) \]
\[ + \beta \sum_{s=1}^{m} \text{Tr} (w_s^T w_s) \]

\[ \Rightarrow \min_w \sum_{s=1}^{m} \text{Tr} (w_s^T (X_s H X_s^T + \beta I_s)^{-1} X_s H f_s) \]

where \( I_s \in \mathbb{R}^{d_s \times d_s} \) is an identity matrix. By setting the derivative w.r.t. \( w_s \) to 0, we have:

\[ 2X_s H X_s^T w_s - 2X_s H f_s + 2\beta w_s = 0 \]

\[ \Rightarrow w_s = (X_s H X_s^T + \beta I_s)^{-1} X_s H f_s = K_s X_s H f_s \]

where \( K_s = (X_s H X_s^T + \beta I_s)^{-1} \).

When optimizing the predicted labels \( f_s \), the problems becomes the following one after we substitute \( w_s \) into the objective function:

\[ \min_{f_s} \sum_{s=1}^{m} \left\| (HX_s^T K_s X_s H f_s - H f_s) \right\|_2^2 + \alpha \sum_{s=1}^{m} \left\| f_s - f \right\|_2^2 \]
\[ + \beta \sum_{s=1}^{m} \left\| K_s X_s H f_s \right\|_2^2 \]

\[ \Rightarrow \min_{f_s} \sum_{s=1}^{m} \text{Tr} \left( f_s^T (HX_s^T K_s X_s H - I_s)^2 f_s \right) \]
\[ + \alpha \sum_{s=1}^{m} \text{Tr} (f_s^T - f) (f_s - f) \]
\[ + \beta \sum_{s=1}^{m} \text{Tr} (f_s^T (K_s X_s H)^2 K_s X_s H f) \]

By setting the derivative w.r.t. \( f_s \) to 0 we obtain:

\[ 2(HX_s^T K_s X_s H - I_n)^2 f_s + 2\alpha f_s - 2\alpha f \]
\[ + 2\beta H X_s^T K_s^2 X_s H f_s = 0 \]

\[ \Rightarrow ((X_s^T K_s X_s - I_n)^2 + \alpha I_n + \beta H X_s^T K_s^2 X_s H) f_s = \alpha f \]

\[ \Rightarrow f_s = \alpha J_s f \]

where \( J_s = ((X_s^T K_s X_s - I_n)^2 + \alpha I_n + \beta H X_s^T K_s^2 X_s H)^{-1} \).

Similarly, by setting the derivative w.r.t. \( b \) to 0 we get:

\[ b^T = \frac{1}{n} \left( I_n f - f \right) X_s^T w \]

To solve the \( \ell_2,1 \)-norm of \( w \), we define a diagonal matrix \( D \) whose diagonal elements \( D_{ii} = \frac{1}{2||w||^2} \). Thus, the problem becomes:

\[ \min_w (HX_s^T w - H f)^T (HX_s^T w - H f) + \beta Tr (w^T D w) \]

By setting the derivative w.r.t. \( w \) to 0, we have:

\[ 2X H X^T w - 2X H f + 2\beta D w = 0 \]

\[ \Rightarrow w = (X H X^T + \beta D)^{-1} X H f = M X H f \]

where \( M = (X H X^T + \beta D)^{-1} \).

Substituting \( w_i, f_i, b_i, w \) and \( b \) into Eq. (5), we arrive at the following problem w.r.t. \( f \):

\[ \min_{f} \sum_{s=1}^{m} \left\| \alpha (H X_s^T K_s X_s H J_s - H J_s) f_s \right\|_2^2 \]
\[ + \left\| H X_s^T M X H f - H f \right\|_2^2 \]
\[ + Tr (f_s^T \sum_{s=1}^{m} \lambda_s^0 \tilde{R}_s F) + \alpha \sum_{s=1}^{m} \left\| (\alpha J_s - I) f_s \right\|_2^2 \]
\[ + \| f - y \|_2^2 + \beta \sum_{s=1}^{m} \left\| \alpha K_s X_s H J_s f_s \right\|_2^2 + \left\| M X H f \right\|_2^2 \]

By setting the derivative w.r.t. \( f \) to 0 we obtain:

\[ f = \alpha \left( \sum_{s=1}^{m} \lambda_s^0 \tilde{R}_s + \alpha \sum_{s=1}^{m} (\alpha J_s - I)^2 \right) \]
\[ +(H X^T M X H - H)^2 + \sum_{s=1}^{m} \lambda_s^0 \tilde{R}_s + \alpha \sum_{s=1}^{m} (\alpha J_s - I)^2 \]
\[ + \alpha I + \beta H X^T M^2 X H \right)^{-1} y \]
After getting $w_i$, $f_i$, $b_i$, $w$, $f$ and $b$, we can solve for the weights $\lambda_s$ as follows.

The problem is equivalent to:

$$\min_{\lambda_s} \text{Tr} \left( f^T \sum_{s=1}^{m} \lambda_s^p \tilde{R}_s f \right)$$

s.t. $\sum_{s=1}^{m} \lambda_s^p = 1, \lambda_s^p \in [0, 1].$ \hspace{1cm} (16)

By using a Lagrange multiplier $\xi$ we convert the above problem to a Lagrange function as:

$$L(\lambda_s, \xi) = \text{Tr} \left( f^T \sum_{s=1}^{m} \lambda_s^p \tilde{R}_s f \right) - \xi \left( \sum_{s=1}^{m} \lambda_s^p - 1 \right).$$ \hspace{1cm} (17)

By setting its derivative w.r.t. $\lambda_s$ and $\xi$ to 0 respectively, we have:

$$\begin{cases} 
 p\lambda_s^{p-1} \text{Tr} \left( f^T \tilde{R}_s f \right) - \xi = 0 \\
 \sum_{s=1}^{m} \lambda_s - 1 = 0
\end{cases} \hspace{1cm} (18)$$

Based on Eq. (18), $\lambda_s$ is attained as:

$$\lambda_s = \left( \frac{1}{\text{Tr} \left( f^T \tilde{R}_s f \right)} \right)^{\frac{1}{p-1}} \sum_{s=1}^{m} \left( \frac{1}{\text{Tr} \left( f^T \tilde{R}_s f \right)} \right)^{\frac{1}{p-1}} \hspace{1cm} (19)$$

Our alternating solution for the objective function in Eq. (5) is summarized as Algorithm 1. It can be proved by the following theorem that the objective function value monotonically decreases in each iteration until convergence using Algorithm 1.

**Theorem 1**: The objective function value shown in Eq. (5) monotonically decreases in each iteration until convergence using the iterative approach in Algorithm 1.

**Proof**: See Appendix A. \hfill $\Box$

The computational complexity is discussed as follows. The complexity of Algorithm 1 is mainly induced by the calculation of the inverse of a few matrices, which is $O \left( \max(d, n)^3 \right)$.

After the classification model is learned, testing is linear w.r.t. the number of testing data $n_{te}$.

## 5 Experiments

In this section, we present the experiments on multimedia event detection with few positive exemplars. The comparison to several algorithms is performed to show the efficacy of our method.

### 5.1 Datasets

We use the TRECVID MED’13 EK10 (MED13) [2] and MED’14 EK10 (MED14) [3] datasets introduced by NIST in the TRECVID challenge. MED13 and MED14 both have 20 multimedia events with 10 events overlapping, resulting in 30 events in total. The 30 events and their examples are displayed in Figure 2. Both datasets are provided with standard training and testing sets. The training sets of MED13 and MED14 are of the same size, comprising 10 positive examples and 5,002 negative examples per event. The testing set consists of 24,957 and 23,953 videos for MED13 and MED14 respectively.

To formulate the semantic representations, we use the TRECVID 2014 semantic indexing task (SIN12) dataset (346 concepts) [4], the Google sports dataset (Sports) (478 concepts) [28], the UCF101 dataset (101 concepts) [43], the Do It Yourself (DIY) dataset (1601 concepts) [26] and the Yahoo Flickr Creative Commons (YFCC) dataset (609 concepts) [1]. Technically, we train concept detectors from these sources and apply them on the MED videos to get the prediction scores as the semantic representation.

### 5.2 Comparison Algorithms

SVM and Ridge Regression (RR) have been widely used by several research groups for TRECVID MED and have shown their robustness [51][33], so we use them as the baseline. Since our focused issue is fast MED, we use linear SVM and linear RR. To be consistent with the implementation of our method,
late fusion is used for these two classifiers w.r.t. the multiple groups of semantic representation.

In addition, we compare our method to several state-of-the-art late fusion algorithms, including GP method [27], LPBoost [20], Adaptive Late Fusion (ALF) [39], and Robust Late Fusion (RLF) [48]. These algorithms similarly leverage the multiple groups of semantic representation.

The parameters of all the comparison algorithms are set as follows. The parameter $p$ in Eq. 5 of B-LSA is set to 10 empirically. Our method has two regularization parameters including $\alpha$ and $\beta$; SVM, RR, GP method, LPBoost, ALF and RLF all have just one regularization parameter. We tune these
Figure 4. MED performance of the comparison algorithms on MED14 dataset.

parameters from \{0.001, 0.01, 0.1, 1, 10, 100, 1000\} uniformly and the best result for each algorithm is reported.

5.3 MED Performance

The MED performance of all the comparison algorithms is presented in Figure 3, Figure 4 and Table 1. It can be seen that our proposed method B-LSA is the most competitive algorithm. Specifically, we have the following observations: 1) B-LSA gains the best performance for 16 events on MED13 dataset and 15 events on MED14 dataset; 2) For the remaining events, B-LSA is the second best performer for 3 events on MED13 dataset and 4 events on MED14 dataset; 3) For
Figure 5. Performance comparison between utilizing semantic interaction and not using it on MED13.

Table 1
Mean Average Precision of all the events on MED13 and MED14 datasets.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>SVM</th>
<th>RR</th>
<th>GP</th>
<th>LPBoost</th>
<th>ALF</th>
<th>RLF</th>
<th>B-LSA</th>
</tr>
</thead>
<tbody>
<tr>
<td>MED13</td>
<td>0.185</td>
<td>0.187</td>
<td>0.189</td>
<td>0.194</td>
<td>0.200</td>
<td>0.209</td>
<td>0.220</td>
</tr>
<tr>
<td>MED14</td>
<td>0.135</td>
<td>0.140</td>
<td>0.141</td>
<td>0.147</td>
<td>0.153</td>
<td>0.162</td>
<td>0.186</td>
</tr>
</tbody>
</table>

many of the events that B-LSA gets the top performance, its advantage over other algorithms is considerably visible; 4) As a state-of-the-art algorithm, ALF is generally the second robust method that effectively exploits the informative cues from the multiple groups of semantic representation; 5) Sophisticated fusion algorithms, i.e., LPBoost, RLF and ALF outperform SVM and RR with late fusion in general, indicating it is beneficial to leverage representations from different sources with certain learning scheme to uncover their disparate contributions to MED; 6) Our method B-LSA achieves the highest score for Mean Average Precision over all the 20 events on both datasets, validating its efficacy. It is worth mentioning that MED is a very challenging task, making it difficult to get substantial improvement by relying on just one technique. The fact that our method gains 5% and 15% relative improvement over the second performer RLF on MED13 and MED14 is encouraging and one could leverage other techniques, e.g., better semantic features to further improve the overall detection accuracy for MED task. We discuss the detection results as follows. Our method treats each representation learned from each source differently. This is more reasonable as different source datasets have different qualities and cover different range of concepts. Therefore, their contribution to event detection is presumably different. By learning weights, our method captures such difference and utilizes it for optimization to get a better fusion scheme. Particularly, we leverage semantic interaction to achieve the goal as it has proved to be effective to uncover data structure.

On the other hand, we simultaneously exert sparsity to the overall semantic representation to suppress the negative influence of the irrelevant or noisy concepts. This component is jointly optimized with the weight learning, thus making them mutually beneficial. Through the joint framework, we obtain more robust detectors together with their weights.

By contrast, SVM and Ridge Regression simply treat these groups of representation equally for late fusion. They overlook the differentiating information, which leads to poorer performance. All the other fusion algorithms have different ways of learning the weights but semantic interaction used by our method seems to be more capable of mining the difference of different sources. Additionally, these fusion algorithms miss the analysis of the overall concept contribution whereas our method considers it simultaneously in the weight learning. Thus, our method performs better than the state-of-the-art fusion methods.

5.4 Impact of Semantic Interaction

In this experiment, we test the usefulness of semantic interaction for MED performance. We reformulate the objective function by removing the semantic interaction term. The new detection results in comparison with the original results are displayed in Figure 5 and Figure 6. We observe that the performance of 29 out of the total 30 events gets worse compared to the original proposed joint model. The MAPs on MED13 and MED14 are 0.191 and 0.152 respectively, which
are also worse than those of the proposed joint model. The results indicate that semantic interaction does contribute to the performance boost.

5.5 Impact of Concept Sparsity

The following experiment aims to study the contribution of the sparse constraint on the overall concept-level to MED performance. By removing this term, we have the results as demonstrated in Figure 7 and Figure 8. Similarly, we also show
the original results. The results show that without the sparse constraint, all the events of MED13 and MED14 get worse results. The MAPs are 0.194 and 0.164 respectively, which are worse than those of the proposed joint model. Therefore, we argue that the concept sparsity constraint is reasonable and is also beneficial for the final detection performance.

5.6 Weights w.r.t. Sources

As claimed, our method is able to learn the weights for the multiple groups of semantic representation resulted from different source datasets. In this part, we present the weights learned from our algorithm in Table 2 and Table 3.

It can be seen that the contribution of each source varies from event to event. For example, the semantic representation learned from SIN dataset is considered the most important for Changing a vehicle tire but the least important for Making a...
Figure 9. Performance comparison between using all groups of semantic representation and leaving out the least useful semantic representation on MED13.

Figure 10. Performance comparison between using all groups of semantic representation and leaving out the least useful semantic representation on MED14.

sandwich; Google sports dataset is considered the most useful source for Parkour but the least useful one for Renovating a home. We also observe that SIN and Google sports datasets are generally evaluated as the most useful sources in contrast with YFCC dataset as the least useful source, and that UCF101 and DIY datasets have respectable significance to the detection.
Figure 11. Performance sensitivity analysis on MED13.

5.7 Leave-One-Out Performance

In the previous experiment, we show that our algorithm is able to treat the multiple groups of semantic representation differently with different weights. Following this scenario, we further conduct an experiment by leaving out the least useful semantic presentation. That is to say, we use four groups of semantic representation in our algorithm for MED. For example, only the semantic representations learned from SIN, Sports, UCF101 and YFCC are cast into our learning method for detecting Birthday party. To show the variation, we compare the resulting detection accuracy to the original one using all the five groups of semantic representation. The
Figure 12. Performance sensitivity analysis on MED14.

According to the results, there is only slight performance variation when the least useful semantic representation is left out. For most events and the average performance, we observe a little performance drop. Interestingly, the accuracy increases for three events, particularly the event Playing fetch for which there is visible improvement. This experiment indicates that 1) using fewer groups of semantic representation does not affect the detection rate considerably; 2) some source data may bring in noise so it is likely to obtain better performance given them not being used. Thus, investigating the usefulness of a source archive before training the concept detectors could be of merit since we could save the computational cost.
5.8 Parameter Sensitivity

We tune the two regularization parameters denoted as $\alpha$ and $\beta$ in Eq. (5) to get the best results. To learn how they affect the detection performance on each event, we conduct an experiment on the parameter sensitivity.

Figure 11 and Figure 12 demonstrate the AP variation w.r.t $\alpha$ and $\beta$ of all the events on MED13 and MED14 respectively. From these two figures we notice that: 1. The detection accuracy varies w.r.t. to different combinations of $\alpha$ and $\beta$. 2. Different events have different sensitivity regarding the parameters. For example, the result does not change much for the event *Parkour* but is quite sensitive for the event *Marriage proposal*. We confer that the impact of different values of the regularization parameters should be related to the trait of the data themselves. Though our method is not very robust to the parameters, in practice one can use cross validation to determine the best parameters during the training to get the best model.

6 Conclusion

In this paper, we have proposed a bi-level semantic representation analyzing framework to harness multiple semantic representations of MED videos learned from different sources more effectively. Regarding the source-level, feature interaction is utilized to capture the semantic interaction within each semantic representation, which is further cast into our framework for weight learning. Regarding the overall concept-level, sparse learning is properly adopted to restrain the negative influence of noisy/irrelevant concepts to the event detection. These two modules are simultaneously optimized with other constraints including prediction consistency and prediction fitting over semantic interaction. As a result, we obtain a family of classification models associated with proper weights for event detection on testing videos. A series of experiments are completed on the TRECVID MED13 EK10 and MED14 EK10 datasets. The comparison with other state-of-the-art algorithms demonstrate the effectiveness of our method.

7 Acknowledgments

This paper was partially supported by the US Department of Defense the U. S. Army Research Office (W911NF-13-1-0277) and the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center contract number D11PC20068. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, ARO, or the U.S. Government.

APPENDIX A

Proof of Theorem 1

**Proof:** According to Algorithm 1, it can be inferred from Eq. (5) that:

$$
\begin{align*}
&\|X^Tw^{t+1} + 1_n(b^{t+1})^T - f^{t+1}\|^2_2 \\
&+ \sum_{s=1}^{m} \left( X_s^Tw^{t+1} + 1_n(b_s^{t+1})^T - f_s^{t+1}\right) \\
&+ Tr \left( (f^{t+1})^T \sum_{s=1}^{m} (\lambda_s^{t+1})^T R_s f^{t+1} \right) \\
&+ \alpha \sum_{s=1}^{m} \|f_s^{t+1} - f^{t+1}\|^2_2 + \|f^{t+1} - y\|^2_2 \\
&+ \beta \left( (w^{t+1})^T Dw^{t+1} + \sum_{s=1}^{m} \|w_s^{t+1}\|^2_2 \right) \\
&\leq \|X^Tw^t + 1_n(b^t)^T - f^t\|^2_2 + \sum_{s=1}^{m} \left( X_s^Tw_s^t + 1_n(b_s^t)^T - f_s^t\right) \\
&+ Tr \left( (f^t)^T \sum_{s=1}^{m} (\lambda_s^t)^T R_s f^t \right) \\
&+ \alpha \sum_{s=1}^{m} \|f_s^t - f^t\|^2_2 + \|f^t - y\|^2_2 \\
&+ \beta \left( (w^t)^T Dw^t + \sum_{s=1}^{m} \|w_s^t\|^2_2 \right)
\end{align*}
$$

Based on the definition of $D$, we have

$$
\begin{align*}
&\|X^Tw^{t+1} + 1_n(b^{t+1})^T - f^{t+1}\|^2_2 \\
&+ \sum_{s=1}^{m} \left( X_s^Tw^{t+1} + 1_n(b_s^{t+1})^T - f_s^{t+1}\right) \\
&+ Tr \left( (f^{t+1})^T \sum_{s=1}^{m} (\lambda_s^{t+1})^T R_s f^{t+1} \right) \\
&+ \alpha \sum_{s=1}^{m} \|f_s^{t+1} - f^{t+1}\|^2_2 + \|f^{t+1} - y\|^2_2 \\
&+ \beta \left( (w^{t+1})^T Dw^{t+1} + \sum_{s=1}^{m} \|w_s^{t+1}\|^2_2 \right) \\
&\leq \|X^Tw^t + 1_n(b^t)^T - f^t\|^2_2 + \sum_{s=1}^{m} \left( X_s^Tw_s^t + 1_n(b_s^t)^T - f_s^t\right) \\
&+ Tr \left( (f^t)^T \sum_{s=1}^{m} (\lambda_s^t)^T R_s f^t \right) \\
&+ \alpha \sum_{s=1}^{m} \|f_s^t - f^t\|^2_2 + \|f^t - y\|^2_2 \\
&+ \beta \left( (w^t)^T Dw^t + \sum_{s=1}^{m} \|w_s^t\|^2_2 \right)
\end{align*}
$$

(21)
The above inequality can be rewritten as:

\[
\|X^T w^{t+1} + 1_n(b^{t+1})^T - f^{t+1}\|_2^2 \\
+ \sum_{s=1}^m \left\|X^T w_s^{t+1} + 1_n(b_s^{t+1})^T - f_s^{t+1}\right\|_2^2 \\
+ Tr \left((f^{t+1})^T \sum_{s=1}^m (X_s^{t+1})^T R_s f^{t+1}\right) \\
+ \alpha \sum_{s=1}^m \left(\|f_s^{t+1} - f^{t+1}\|_2^2 + \|f^{t+1} - y\|_2^2\right) + \beta \sum_i \|w_i^{t+1}\|_2^2 \\
\leq \|X^T w^t + 1_n(b^{t})^T - f^t\|_2^2 \\
+ \sum_{s=1}^m \left\|X^T w_s^t + 1_n(b_s^{t})^T - f_s^t\right\|_2^2 \\
+ Tr \left((f^{t})^T \sum_{s=1}^m (X_s^t)^T R_s f^t\right) \\
+ \alpha \sum_{s=1}^m \left(\|f_s^t - f^t\|_2^2 + \|f^t - y\|_2^2\right) + \beta \|w_i^t\|_2^2 \\
\] (22)

According to the definition of \(\ell_{2,1}\)-norm, it becomes:

\[
\|X^T w^{t+1} + 1_n(b^{t+1})^T - f^{t+1}\|_2^2 \\
+ \sum_{s=1}^m \left\|X^T w_s^{t+1} + 1_n(b_s^{t+1})^T - f_s^{t+1}\right\|_2^2 \\
+ Tr \left((f^{t+1})^T \sum_{s=1}^m (X_s^{t+1})^T R_s f^{t+1}\right) \\
+ \alpha \sum_{s=1}^m \left(\|f_s^{t+1} - f^{t+1}\|_2^2 + \|f^{t+1} - y\|_2^2\right) + \beta \|w_i^{t+1}\|_2^2 \\
\leq \|X^T w^t + 1_n(b^{t})^T - f^t\|_2^2 \\
+ \sum_{s=1}^m \left\|X^T w_s^t + 1_n(b_s^{t})^T - f_s^t\right\|_2^2 \\
+ Tr \left((f^{t})^T \sum_{s=1}^m (X_s^t)^T R_s f^t\right) + \alpha \sum_{s=1}^m \left(\|f_s^t - f^t\|_2^2 + \|f^t - y\|_2^2\right) + \beta \|w_i^t\|_2^2 \\
\] (25)

which indicates that the objective function value of Eq. (5) monotonically decreases until converging to the optimal \(w\) through the proposed approach in Algorithm 1.

\[\square\]

\section*{References}


Xiaojun Chang is a Ph.D. student at University of Technology Sydney. His research interests include machine learning, data mining and computer vision. His publications appear in proceedings of prestigious international conferences like ICML, ACM MM, AAAI, IJCAI and etc.

Zhizhang Ma received the Ph.D. in computer science from University of Trento, Trento, Italy, in 2013. He is now a Postdoctoral Research Associate with the School of Computer Science, Carnegie Mellon University, Pittsburgh, PA. His research interest is mainly on machine learning and its applications to multimedia analysis and computer vision.
Yi Yang received the Ph.D. degree in computer science from Zhejiang University. He was a Post-Doctoral Research Fellow with the School of Computer Science, Carnegie Mellon University, from 2011 to 2013. He is currently a Senior Lecturer with the Centre for Quantum Computation and Intelligent Systems, University of Technology at Sydney, Sydney. His research interests include multimedia and computer vision.

Zhiqiang Zeng received the Ph.D. in computer science from Zhejiang University.
He is an associate professor at Xiamen University of Technology. His main research interest includes statistics learning theory and pattern recognition.

Alexander G. Hauptmann received the B.A. and M.A. degrees in psychology from Johns Hopkins University, Baltimore, MD, the degree in computer science from the Technische Universität Berlin, Berlin, Germany, in 1984, and the Ph.D. degree in computer science from Carnegie Mellon University (CMU), Pittsburgh, PA, in 1991.

He is currently with the faculty of the Department of Computer Science and the Language Technologies Institute, CMU. His research interests include several different areas: man-machine communication, natural language processing, speech understanding and synthesis, video analysis, and machine learning. From 1984 to 1994, he worked on speech and machine translation, when he joined the Informedia project for digital video analysis and retrieval, and led the development and evaluation of news-on-demand applications.