Multimedia Event Detection Using A Classifier-specific Intermediate Representation

Zhigang Ma, Yi Yang, Nicu Sebe, Kai Zheng, and Alexander G. Hauptmann

Abstract—Multimedia event detection (MED) plays an important role in many applications such as video indexing and retrieval. Current event detection works mainly focus on sports and news event detection or abnormality detection in surveillance videos. Differently, our research aims to detect more complicated and generic events within a longer video sequence. In the past, researchers have proposed using intermediate concept classifiers with concept lexica to help understand the videos. Yet it is difficult to judge how many and what concepts would be sufficient for the particular video analysis task. Additionally, obtaining robust semantic concept classifiers requires a large number of positive training examples, which in turn has high human annotation cost. In this paper, we propose an approach that exploits the external concepts-based videos and event-based videos simultaneously to learn an intermediate representation from video features. Our algorithm integrates the classifier inference and latent intermediate representation into a joint framework. The joint optimization of the intermediate representation and the classifier makes them mutually beneficial and reciprocal. Effectively, the intermediate representation and the classifier are tightly correlated. The classifier dependent intermediate representation not only accurately reflects the task semantics but is also more suitable for the specific classifier. Thus we have created a discriminative semantic analysis framework based on a tightly coupled intermediate representation. Extensive experiments on multimedia event detection using real-world videos demonstrate the effectiveness of the proposed approach.

Index Terms—Intermediate Representation, Multimedia Event Detection, p-norm.

1 Introduction

Research on video indexing and retrieval has long been faced with the challenge of semantic gap between low-level features and high-level semantic content description of videos [1][2]. To bridge the semantic gap, various approaches have been proposed to help analyze the semantic content of videos, either at concept level or at event level.

According to [3], a "concept" means an abstract or general idea inferred from specific instances of objects, scenes and actions such as *fish*, *outdoor* and *boxing*. Concepts are lower level descriptions of multimedia data which usually can be inferred with a single image or a few video frames. An "event" refers to an observable occurrence that interests users. Compared with concepts, events are higher level descriptions of multimedia data. A meaningful event builds upon many concepts and is unlikely to be inferred with a single image or a few video frames. For example, the event *landing a fish* includes many concepts such as *people*, *fish*, *fishing rod* together with the action *landing*, and it usually happens in a longer video sequence. We cannot tell if it is a *landing a fish* event if we only see a person sitting on a boat in one image or a few frames.

- *Z. Ma and N. Sebe are with the Department of Information Engineering and Computer Science, University of Trento. E-mail: ma, sebe@disi.unitn.it*
- *Y. Yang and A. Hauptmann are with the School of Computer Science, Carnegie Mellon University. E-mail: yiyang, alex@cs.cmu.edu*
- *K. Zheng is with the School of Information Technology & Electrical Engineering, the University of Queensland. E-mail: kevinz@itee.uq.edu.au*

Annotation and detection are two different topics of both concept and event analysis [3]. Multimedia annotation, also known as recognition, aims to associate a datum with one or multiple semantic labels (tags) [3]. Many approaches have been proposed to improve annotation accuracy for both images and videos [6]. Detection identifies the occurrence of a class of interest in a large pool of data. In contrast with annotation for which both the training and testing data are from a fixed number of classes, the training and testing data in detection can be from an infinite number of classes [3]. Hence, detection is a more challenging problem.

The TREC Video Retrieval Evaluation (TRECVID) community has notably contributed to the research of video concept or event detection [7][8][9]. In the field of multimedia, many other works have also focused on *concept detection*, e.g., [4][10][11]. However, the research on video *event detection* is still in its infancy. Most existing research on event detection is limited to the sport events, news events, events with repetitive patterns like *running* or unusual events in surveillance videos [12][5] [13][14]. The "Event detection in Internet multimedia (MED)[1]" launched by TRECVID aims to encourage new technologies for detecting more complicated events, *e.g.*, *feeding an animal*. Ma *et al.* have made the first attempt on Ad Hoc detection of this type of events, for which only 10 positive example are available for training [3]. For this kind of events, there are huge intra-class variation. For example, an event "feeding an animal" can be either feeding a cat at home with cat food in a small container, or feeding a horse in a farm with a bundle of grass. Besides, they are usually characterized by long video sequences, which necessitates the exploration of all the sequences for analysis.

1. http://www.nist.gov/itl/iad/mig/med11.cfm

Recent research has shown that the performance of multimedia semantic analysis can be improved through proper machine learning approaches [16][17][18]. Therefore, it is reasonable to leverage good low-level features as well as effective machine learning algorithms on video data for MED. We propose a new algorithm for MED, which is extended from our previous work [19]. Our method has the following attributes:

1) Our algorithm learns an intermediate representation of videos by exploiting the *target videos* and *external video* archives together. In this paper, the target videos are the videos depicting the event to be detected. The external videos are the auxiliary labeled video archives that are used to help learn the intermediate representation. The intermediate representation is a compact vector representation derived from the Bag-of-Words features of the videos through a transformation, during which the discriminative information is encoded.

2) Our algorithm integrates representation inference and classifier training into a joint framework. In this way, the intermediate representation is tightly coupled with the loss function used for the classifier.

3) A robust loss function is used in our objective function, making the performance more robust to outliers.

We name our method Semantic Analysis via Intermediate Representation (SAIR). The intermediate representation is dependent on the classifier while the classifier training benefits from the representation. The mutual benefit and reciprocality between the intermediate representation and the classifier endows the classification framework good capability for multimedia event detection.

2 Related Work

In this section, we briefly review some related works, which cover multimedia representation and semantics understanding.

2.1 Multimedia Low-level Feature Representation

A common approach for low-level feature representation is to extract the key frames of videos and then generate features based on these frames. For example, traditional features include Color Correlogram, Edge Direction Histogram, Wavelet Texture, *etc.* Newly designed features, *e.g.*, SIFT draw more research interest for their discriminating capability [20]. Some other features can capture the spatial-temporal information, *e.g.*, STIP feature [21] and MoSIFT feature [22], and have shown promising performance in video semantic analysis.

Apart from visual features, some other modalities, which provide different yet complementary information, can also be used to represent videos. For example, textual representation based on Automatic Speech Recognition (ASR) and Optical Character Recognition (OCR), and auditory features based on Mel-frequency Cepstral Coefficients (MFCC) have also been frequently used to represent videos [23].

2.2 Learning to Refine Multimedia Representation

Multimedia representation refinement aims to obtain a more compact as well as accurate feature representation of multimedia data [15][5][24][17][25]. Shyu *et al.* propose a subspace based data mining framework for video concept/event detection [5]. To exploit the semantic relatedness among multiple modalities, Yang *et al.* propose a manifold learning based algorithm to infer a unified representation of different media types for cross media retrieval [24]. Based on users' feedbacks, a long term relevance feedback algorithm is proposed in [17] to refine the multimedia representation for better retrieval performance. In [25], a sparse projection method is proposed to infer a sparse representation for videos, by which the efficiency of video classification is improved. These research efforts have shown that multimedia data can be refined by proper machine learning algorithms, thus resulting in better performance for multimedia analysis. However, in most of these works, the refinement and the classifier training are independent from each other. As it is uncertain which classifiers benefit the most from these refinement algorithms, the performance improvement could be limited. Instead, we propose an integrated framework which learns a refined representation and a classifier jointly. As the refined representation is correlated with the loss function used in the classifier, the classifier dependent intermediate representation not only accurately reflects the task semantics but is also more suitable for the specific classifier, thus resulting in boosted classification accuracy.

2.3 Concepts-based Representation

Recently, some researchers suggest using concepts-based representation for video semantic understanding. A number of researchers have been building a variety of semantic concept detectors, such as those related to people (face, anchor), acoustic (speech, music), genre (weather, financial, sports), scene, *etc.* [1], and a series of concept lexica have been established, *e.g.*, LSCOM [26] and MediaMill [4]. 346 concepts have been defined for the TRECVID 2011 semantic indexing task. With these annotation corpora, different concept detectors can be trained. Therefore, videos can be represented by the concept detection results of those detectors [27]. If sufficient concept detectors are properly trained and appropriately applied, the concepts-based representation of videos, which is a set of textual descriptors, is more capable of reflecting video semantics. However, such approach is still confronted with some problems. First, it requires many labeled data to train intermediate concept classifiers, which costs much human labor. For example, while the full LSCOM set contains over 2600 concepts, many of them are unannotated or contain no positive instances [26]. Second, only concept-based archives have been used to infer representation so far. In recent years, several event-based video archives have been presented in the community. Effective usage of these event-based videos could be another potential solution for improving multimedia event detection.

3 The Proposed Algorithm

In this section, our algorithm is presented in details followed by an algorithm for solving the objective function. *Classifier-specific* in our method means being tightly coupled with the particular loss function used by the classifier.

3.1 Learning An Intermediate Representation

We first illustrate the traditional approach of concepts-based representation for multimedia analysis. Then we formulate our method which goes beyond the traditional approach.

3.1.1 Traditional Approach

Suppose there are n example videos, whose low-level features are $\{x_1, ..., x_n\}$. Here $x_i \in \mathbb{R}^d$ denotes the low-level feature of the video and d is the dimension of the feature. x_i $(1 \leq i \leq n)$ can be either positive or negative examples. Let y_i be the label of x_i, indicating whether the video x_i is a positive one. A general approach to train a multimedia event detector f can be formulated as minimizing the following objective function

$$\min_f \sum_{i=1}^{n} \ell\left(f(x_i), y_i\right) + \alpha\Omega(f), \tag{1}$$

where $\ell(\cdot,\cdot)$ is a loss function and $\Omega(f)$ is a regularization function on f with α as a regularization parameter. Once f is obtained, we can use it for event detection. Clearly, there are three main components needed to be properly designed, which are the feature representation x_i, the loss function $\ell(\cdot,\cdot)$, and the regularization function $\Omega(\cdot)$.

Using the concepts-based representation as in [27][1] for multimedia event detection, we need another m annotated videos $\{x_{n+1}, ..., x_{n+m}\}$ from c classes with groundtruth labels $\{y_{n+1}, ..., y_{n+m}\}$. For the k-th class there are m_k positive examples. The videos $\{x_{n+1}, ..., x_{n+m}\}$ are used to pre-train c classifiers $g_k|_{k=1}^c$, one for each intermediate concept. For each training or testing video $x_i (1 \leq i \leq n)$, the classifiers $g_k|_{k=1}^c$ are applied to detect the intermediate concepts. In this way, $x_i (1 \leq i \leq n)$ is represented by a c dimensional vector, with each dimension corresponding to an intermediate concept. More specifically, the following two steps are taken. In the first step, c classifiers $\{g_1, ..., g_c\}$ are trained by minimizing the following objective function

$$\min_{g_1,...,g_c} \sum_{k=1}^{c} \sum_{j=1}^{m} \widetilde{\ell}\left(g_k(x_{n+j}), y_{n+j}\right) + \alpha\widetilde{\Omega}(g_k), \tag{2}$$

where $\widetilde{\ell}(\cdot,\cdot)$ and $\widetilde{\Omega}(f)$ are the loss function and the regularization function respectively and α is a parameter. Once the c classifiers $\{g_1, ..., g_c\}$ are obtained, we convert the original feature representation $x_i (1 \leq i \leq n)$ to the concepts-based representation $z_i = [z_{1i}, ..., z_{ci}] \in \mathbb{R}^c$ by $z_{ki} = g_k(x_i)$ $(1 \leq k \leq c)$. In the second step, the event detector f can be trained based on the new representation z_i $(1 \leq i \leq n)$ in the same way of (1), *i.e.*,

$$\begin{aligned} &\min_f \sum_{i=1}^{n} \ell\left(f(z_i), y_i\right) + \alpha\Omega(f) \\ \Rightarrow &\min_f \sum_{i=1}^{n} \ell\left(f(g(x_i)), y_i\right) + \alpha\Omega(f), \end{aligned} \tag{3}$$

where $g(x_i) = [g_1(x_i), ..., g_c(x_i)]$. For each testing video x_{te}, the decision score s_{te} indicating whether the event occurs in the video x_{te} is given by

$$s_{te} = f\left(g(x_{te})\right). \tag{4}$$

Although the traditional concepts-based representation [1][27] is expected to be more precise than low-level features, this kind of approach suffers from some practical problems in implementation. First, it is time-consuming to find and annotate a large amount of positive examples to train many concept classifiers. Second, the number of concepts is limited and it remains unclear how many concepts (and what concepts as well) would be sufficient for some applications, *e.g.*, multimedia event detection. Third, the pre-trained concept classifiers are yet to be sufficiently reliable. Fourth, given a particular event to detect, only some concepts are discriminative while others are comparatively useless or even noisy. Taking "landing a fish" event as an example, some concepts like "fish" and "boat" are very discriminative, while "clouds" and "face" are less informative. It is a nontrivial task to define the ontology for different events, which are dynamic and diverse.

3.1.2 Joint Learning of Classifier and Representation with External Videos

In the traditional way of multimedia event detection using concepts-based representation, the concept classifiers $g_k|_{k=1}^c$ and multimedia event detector f are trained individually, as shown in (2) and (3). There is no guarantee, however, that the two are tightly correlated. Besides, training a large number of $g_k|_{k=1}^c$ is time consuming, while it remains unclear how large c should be. A question then comes up: Can we learn an intermediate representation closely related to a particular multimedia event, and the event detector without requiring many pre-labeled data? As demonstrated in [3], the classifier of external concepts-based videos and the event detector have shared components. Exploiting such information is beneficial for multimedia event detection. Different from [3],we assume that the external concepts-based videos and the event-based videos have a common intermediate representation. Specifically, we propose to simultaneously learn f and an intermediate representation built upon $g_k|_{k=1}^c$ from the external videos and g_{c+1}, g_{c+2} from the positive and negative examples of the particular event to be detected:

$$\min_{f,\{g_1,...,g_{c+2}\}} \sum_{i=1}^{n+m} \ell\left(f([g_1(x_i), ..., g_{c+2}(x_i)]), y_i\right) + \alpha\Omega(f), \tag{5}$$

where $x_i (1 \leq i \leq m+n)$ is either a positive or negative example of a particular event, or an example of external videos used to help learn the intermediate representation. In (5) the classifier and the intermediate representation are jointly optimized, which explicitly guarantees that the two are correlated. Inspired by [28], we define $f(x_i)$ and $g(x_i)$ as follows:

$$f(g(x_i)) = W^T g(x_i), \tag{6}$$

$$g(x_i) = [\theta_1^T x_i, ..., \theta_{c+2}^T x_i] = \Theta^T x_i. \tag{7}$$

Then we rewrite (5) as

$$\min_{W,\Theta} \sum_{i=1}^{n+m} \ell\left(W^T(\Theta^T x_i), y_i\right) + \alpha\|W\|_F^2. \tag{8}$$

In our previous work [19], we used the $\ell_{2,1}$-norm based loss function and obtained good performance for multimedia understanding. In this extension, we apply the $\ell_{2,p}$-norm ($0 < p < 2$) based loss function as we can adjust the value of p to search for the optimal loss. In this way, our previous work is a special case of this new formula. For an arbitrary matrix $A \in \mathbb{R}^{d\times c}$, $\|A\|_{2,p}$ is defined as:

$$\|A\|_{2,p} = \left(\sum_{i=1}^{d}(\sum_{j=1}^{c}|A_{ij}|)^{\frac{p}{2}}\right)^{\frac{1}{p}}. \tag{9}$$

We propose our objective function as:

$$\begin{aligned}&\min_{W,\Theta,b}\left\|X\Theta W + 1_{n+m}b^T - Y\right\|_{2,p} + \alpha\|W\|_F^2.\\ &s.t. \Theta^T\Theta = I\end{aligned} \tag{10}$$

In (10), $X = [x_1, x_2, ..., x_n, x_{n+1}, ..., x_{n+m}] \in \mathbb{R}^{(n+m)\times d}$ is the data matrix including the positive and negative examples $x_1, x_2, ..., x_n$ of a particular event together with the external videos $x_{n+1}, ..., x_{n+m}$. $Y = [y_1, y_2, ..., y_n, y_{n+1}, ..., y_{n+m}] \in \mathbb{R}^{(n+m)\times(c+2)}$ indicate their labels. Note that the external videos have c classes and the positive and negative examples for an event are treated as two classes so we have $c+2$ classes in total. $1_{n+m} \in \mathbb{R}^{n+m}$ is a column vector with all ones and $b \in \mathbb{R}^{c+2}$ is the bias. The bias is added for unbalanced data but we can preprocess the data by centering them. The orthogonal constraint $\Theta^T\Theta = I$ is added for two considerations: 1) to avoid arbitrary scaling of the intermediate representation; 2) to preserve as much information as possible [29]. Suppose the data are centered, (10) becomes:

$$\begin{aligned}&\min_{W,\Theta}\|X\Theta W - Y\|_{2,p} + \alpha\|W\|_F^2.\\ &s.t. \Theta^T\Theta = I\end{aligned} \tag{11}$$

Note that although (11) looks similar to the objective function in [28], our proposed method is different from that of [28]. The primary difference is that the motivation of [28] is to address multi-label classification whereas ours manages to learn an intermediate representation coupled with the specific loss function. When the loss function changes, the intermediate representation, *i.e.*, Θ changes accordingly. Another difference is that we use an $\ell_{2,p}$-norm based loss function which is more robust.

Next, we discuss how the proposed approach tackles the four problems below (4) that are faced by the traditional concepts-based representation methods. First, to obtain good concept classifiers, it usually requires a large amount of labeled training data. Our method, however, does not directly use the concept classifiers but learns an intermediate representation so not many data are required, which is also validated by our experiment. To detect the event *feeding an animal*, traditional methods would train the concept classifier of "animal." However, it is hard to know what concepts else can be useful. If the event happens indoor, concepts such as "floor" would help. If the event happens outdoor, "grass land" is more informative. It is tricky to decide what concepts should be trained in advance. Differently, our method learns an intermediate representation, which does not directly use the pre-defined concept classifiers to perform MED. As can be seen, our method jointly optimizes the loss function and the intermediate representation. In this case, the loss function is optimized for *feeding an animal*. As this learning process is coupled with the detector, it is able to adjust $g(x)$ for the event. When the event is changed, X and Y in (11) will also be different. Consequently, the optimal Θ will be different, which means that different intermediate representations are learned for different events. However, traditional approach uses the same concept detection results for different events, and therefore the selection of concepts turns to a critical problem for the traditional concepts-based representation. Third, traditional methods directly use the output from trained concept classifiers as input for event detection. If the output of the pre-trained classifiers is not reliable, the performance of MED degrades. Differently, our method learns a discriminative intermediate representation, which dose not directly use the output of concept clarifiers as input. Fourth, if we use traditional pre-trained concept classifiers for event detection, we have to decide in advance what concept classifiers to use. In contrast, our method learns g and f jointly with the assumption that concept classifiers and event detector have an intermediate representation. Consequently, we do not need to select the concepts for a particular event.

Algorithm 1: The SAIR algorithm.

Input:
The training data X and the label matrix Y;
Parameter α.

Output:
Converged Θ and W.

1: Set $t = 0$ and initialize Θ_0, W_0 randomly;
2: **repeat**
Compute $[z_t^1, ..., z_t^{n+m}]^T = X\Theta_t W_t - Y$;
Compute the diagonal matrix $\tilde{D}_t$ as:

$$\tilde{D}_t = \begin{bmatrix} \frac{1}{\frac{2}{p}\|z_t^1\|_2^{2-p}} & & \\ & \cdots & \\ & & \frac{1}{\frac{2}{p}\left\|z_t^{n+m}\right\|_2^{2-p}} \end{bmatrix};$$

Compute $U_t = X^T\tilde{D}_t X + \alpha I$;
Compute $V_t = X^T\tilde{D}_t Y Y^T\tilde{D}_t X$;
Obtain Θ_{t+1} by the eigen-decomposition of $U_t^{-1}V_t$;
Compute $A_t = \Theta_t^T X^T\tilde{D}_t X\Theta_t + \alpha I$;
Update W_{t+1} as $W_{t+1} = A_t^{-1}\Theta_t^T X^T\tilde{D}_t Y$;
$t = t + 1$.
until *Convergence*;
3: Return Θ and W.

3.2 Solution

The $\ell_{2,p}$-norm in our framework is non-smooth which makes (11) difficult to solve. To deal with this problem, we propose the following solution. By denoting $X\Theta W - Y = [z^1, ..., z^{n+m}]^T$, the objective of (11) is equivalent to:

$$\begin{aligned}&\min_{W,\Theta} Tr\left((X\Theta W - Y)^T\tilde{D}(X\Theta W - Y)\right) + \alpha\|W\|_F^2,\\ &s.t. \quad \Theta^T\Theta = I\end{aligned} \tag{12}$$

where $\tilde{D}$ is a matrix with its diagonal elements $\tilde{D}_{ii} = \frac{1}{\frac{2}{p}\|z^i\|_2^{2-p}}$. By setting the derivative *w.r.t.* W to 0, we have:

$$W = A^{-1}\Theta^T X^T \tilde{D} Y, \tag{13}$$

where $A = \Theta^T X^T \tilde{D} X \Theta + \alpha I$ and I is an identity matrix. The above procedure needs to calculate the inverse of A. $A = \Theta^T X^T \tilde{D} X \Theta + \alpha I = (X\Theta)^T \tilde{D}(X\Theta) + \alpha I$. As D is semi-positive, $(X\Theta)^T \tilde{D}(X\Theta)$ is semi-positive. I is positive definite. Thus, A is non-singular and invertible. Substituting (13) into (12), it becomes:

$$\begin{aligned} &\min_{\Theta} Tr\left(Y^T \tilde{D} X \Theta A^{-1}(\Theta^T X^T \tilde{D} X \Theta - 2A + \alpha I)\right. \\ &\left. A^{-1}\Theta^T X^T \tilde{D} Y\right) \\ &\qquad s.t. \quad \Theta^T \Theta = I \end{aligned} \tag{14}$$

As $A = \Theta^T X^T \tilde{D} X \Theta + \alpha I$, (14) becomes:

$$\begin{aligned} &\max_{\Theta} Tr\left((\Theta^T U \Theta)^{-1}\Theta^T V \Theta\right), \\ &\qquad s.t. \quad \Theta^T \Theta = I \end{aligned} \tag{15}$$

where $U = X^T \tilde{D} X + \alpha I$ and $V = X^T \tilde{D} Y Y^T \tilde{D} X$.

The objective function of (15) can be readily solved by the eigen-decomposition of $U^{-1}V$. However, the solving of Θ requires the input of $\tilde{D}$ that is related to W, so it is not handy to get Θ and W. Therefore, we propose an iterative approach demonstrated in Algorithm 1. It can be proved that the objective function value shown in (11) monotonically decreases in each iteration until convergence using the iterative approach in Algorithm 1. The complexity of calculating the inverse of a few matrices is $\mathcal{O}(d^3)$. To obtain Θ, we need to conduct eigen-decomposition of $U^{-1}V$, which is also $\mathcal{O}(d^3)$ in complexity.

3.3 Nonlinear SAIR

As nonlinear classifiers generally have better performance than linear ones for event detection [23], we extend our algorithm SAIR to a nonlinear classifier by utilizing kernel tricks. Assuming that there is a transformation function $\phi : \mathbb{R}^d \to \mathcal{H}$. Then, the objective function of the nonlinear SAIR can be written as:

$$\begin{aligned} &\min_{W,\phi(\Theta)} \|\phi(X)\phi(\Theta)W - Y\|_{2,p} + \alpha \|W\|_F^2, \\ &s.t. \quad \phi(\Theta)^T \phi(\Theta) = I \end{aligned} \tag{16}$$

It has been proved in [31] that if we map the data into a Hilbert space $\mathcal{H}$ by Kernelized Principal Component Analysis (KPCA) [30], (16) can be solved in a similar way as (11) using the representations in $\mathcal{H}$.

4 Experiments

In this section, we present the experimental results. We use the nonlinear SAIR with χ^2 kernel. Our method is compared to the following algorithms: AdaBoost, TaylorBoost [32], SVM, Linear Discriminant Analysis (LDA) [33] followed by ridge regression and Semantic Concept Representation (SCR). For SCR, we use the existing concept-based video corpus to learn the representation of the event-based videos. Then SVM with χ^2 kernel is applied for classification.

4.1 Datasets

We use the TRECVID MED 2011 (MED11)[2] development set in our experiments, which includes 15 events. We perform event detection for these 15 events.

Another two video sets, *i.e.*, the TRECVID MED 2010 (MED10)[3] and the development set from TRECVID 2011 semantic indexing task are used as external video sources. We use them to help learn the intermediate representation for MED11. MED10 includes 3 events. The video set for semantic indexing task covers 346 concepts. We used 65 concepts suggested by [34]. These concepts are related to human, environment and object. For convenience, we denote the resulting dataset as Semantic Indexing dataset(SIN11). Recall that in (11) $Y \in \mathbb{R}^{(n+m)\times(c+2)}$ where $c = 3+65 = 68$ in our setting. According to the task definition from NIST, each event is detected independently. In our experiments, there are 15 individual detection tasks.

4.2 Setup

The training data comprise three parts. The first part consists of 100 positive examples and 500 negative examples randomly selected from MED11. The second part includes 309 positive examples from MED10. The third part is SIN11 which has 2529 video frames. The remaining videos in MED11 are our testing data.

We use a 4096 dimension Bag-of-Words feature to represent each video using SIFT, CSIFT [35] and MoSIFT separately. The three feature types are further concatenated. We ran our program on the Carnegie Mellon University Parallel Data Lab cluster, which contains 300 cores, to extract features and perform the bag-of-words mapping. The parameters of all algorithms in our experiments are tuned by a "grid-search" strategy from $\{10^{-3}, 10^{-2}, \cdots, 10^2, 10^3\}$. We use two evaluation metrics. The first one, Minimum NDC (MinNDC) [36], is defined as follows:

$$\begin{aligned} &MinNDC(S,E) \\ &= \frac{C_M P_M(S,E) P_T + C_{FA} P_{FA}(S,E)\,(1 - P_{FA}(S,E))}{MINUMUM\,(C_M P_T, C_M(1-P_T))}, \end{aligned} \tag{17}$$

where $P_M(S,E)$ is the missed detection probability for system S, event E while $P_{FA}(S,E)$ is the false alarm probability for system S, event E. $CM = 80$ is the cost for missed detection, $C_{FA} = 1$ is the cost for false alarm and $PT = 0.001$. Lower MinNDC indicates better detection performance. The second one is Average Precision (AP). Higher AP indicates better performance.

4.3 MED Results

The MED results are displayed in Table 1 using the two evaluation metrics. It can be seen that our method SAIR is consistently competitive compared with other methods. Zooming into details, we have the following observations: 1) In terms of MinNDC, SAIR gains the best performance for 9 events and the second best performance for another 5 events.

2. http://www.nist.gov/itl/iad/mig/med11.cfm
3. http://nist.gov/itl/iad/mig/med10.cfm

Table 1
MED performance comparison. Note that LOWER MinNDC / HIGHER AP indicates BETTER performance. The best results are highlighted in bold.

Event Description	Evaluation Metric	AdaBoost	TaylorBoost	SVM	LDA	SCR	SAIR
Attempting a board trick	MinNDC	1.218	0.995	0.826	0.998	**0.742**	0.775
	AP	0.086	0.094	0.225	0.131	**0.274**	0.248
Feeding an animal	MinNDC	1.343	1.001	**0.963**	1.001	0.981	0.964
	AP	0.037	0.043	0.087	0.045	0.079	**0.089**
Landing a fish	MinNDC	1.119	0.932	0.665	0.938	0.704	**0.626**
	AP	0.065	0.097	0.260	0.103	0.234	**0.281**
Wedding ceremony	MinNDC	1.015	1.001	0.466	1.001	0.582	**0.441**
	AP	0.084	0.067	0.483	0.073	0.322	**0.493**
Working on a woodworking project	MinNDC	1.203	1.001	0.726	1.001	0.940	**0.711**
	AP	0.055	0.046	**0.294**	0.096	0.091	0.283
Birthday party	MinNDC	1.211	1.001	0.885	1.001	0.939	**0.882**
	AP	0.030	0.019	**0.079**	0.021	0.051	0.076
Changing a vehicle tire	MinNDC	1.187	1.001	0.670	1.001	0.862	**0.636**
	AP	0.006	0.006	0.023	0.006	0.013	**0.030**
Flash mob gathering	MinNDC	1.139	1.001	0.629	1.001	**0.509**	0.568
	AP	0.050	0.042	0.198	0.059	**0.291**	0.228
Getting a vehicle unstuck	MinNDC	1.031	0.902	0.802	0.970	**0.586**	0.711
	AP	0.019	0.027	0.051	0.018	**0.107**	0.083
Grooming an animal	MinNDC	1.317	1.001	0.856	0.925	**0.814**	0.856
	AP	0.006	0.013	0.046	0.025	**0.056**	0.047
Making a sandwich	MinNDC	1.355	1.001	**0.821**	1.001	0.843	0.858
	AP	0.008	0.009	**0.034**	0.010	0.029	0.030
Parade	MinNDC	1.091	0.991	0.654	1.001	0.712	**0.632**
	AP	0.035	0.028	0.093	0.019	0.083	**0.108**
Parkour	MinNDC	1.156	0.955	0.570	1.001	0.566	**0.449**
	AP	0.014	0.005	0.047	0.009	0.050	**0.055**
Repairing an appliance	MinNDC	0.971	1.001	0.550	0.822	0.664	**0.508**
	AP	0.027	0.018	0.102	0.029	0.056	**0.109**
Working on a sewing project	MinNDC	1.188	1.001	0.706	0.974	0.833	**0.612**
	AP	0.012	0.008	0.037	0.016	0.027	**0.054**
Average	MinNDC	1.163	0.986	0.719	0.976	0.752	**0.682**
	AP	0.035	0.035	0.137	0.044	0.118	**0.148**

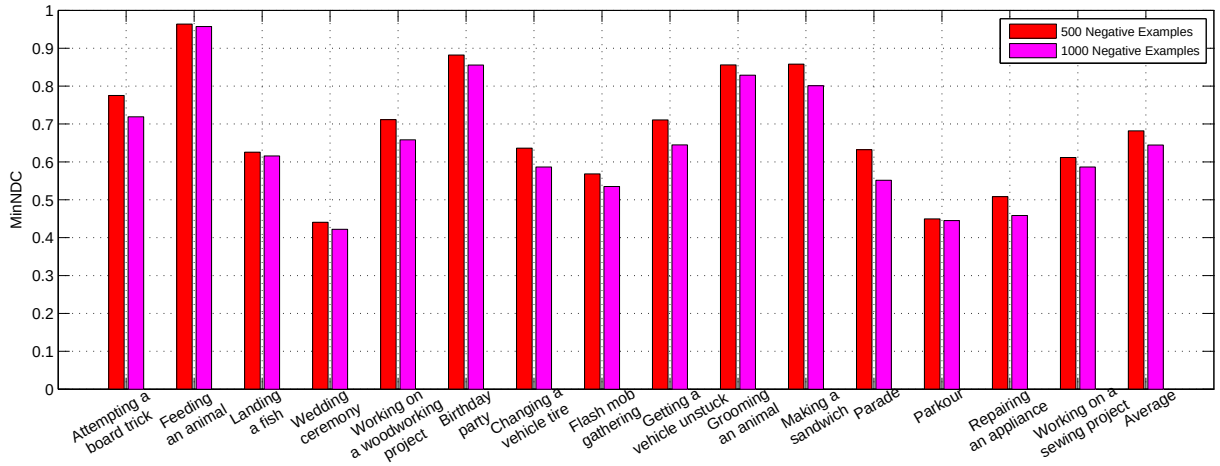
Table 2
Performance comparison between using 30 concepts and using 65 concepts from SIN11.

Event Description	Evaluation Metric	SCR(30C)	SCR(65C)	SAIR(30C)	SAIR(65C)
Attempting a board trick	MinNDC	0.811	0.742	0.764	0.775
	AP	0.215	0.274	0.246	0.248
Feeding an animal	MinNDC	0.976	0.981	0.961	0.964
	AP	0.071	0.079	0.091	0.089
Landing a fish	MinNDC	0.722	0.704	0.625	0.626
	AP	0.214	0.234	0.286	0.281

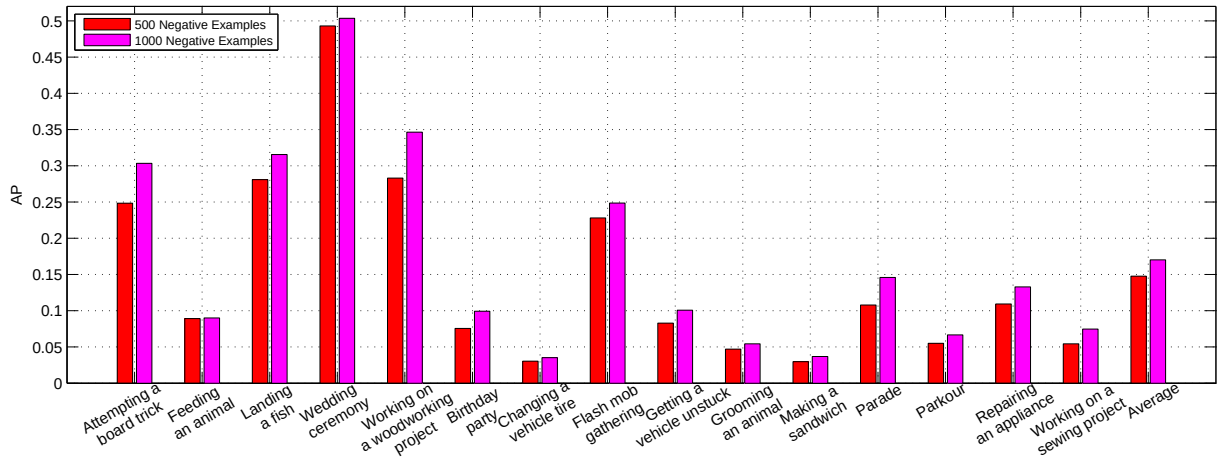
SAIR outperforms all other methods for the average accuracy over all the 15 events. 2) In terms of AP, SAIR is the best method for 8 events and the second best one for the other 7 events. SAIR obtains the top performance for the average accuracy over all the 15 events. Notably, it outperforms the runner-up SVM by 8%. 3) SVM and SCR have varying degree of success for some events. However, when considering the overall performance, they are not as consistently robust as SAIR. 4) As a linear approach, LDA has weak performance. Hence, it is preferable to use kernel methods. The better performance of SAIR indicate that leveraging other concept-based and/or event-based videos is beneficial for multimedia event detection.

4.4 Performance *w.r.t.* Fewer Concepts

To study whether the number of concepts selected affects the MED performance, we conduct an experiment by reducing the 65 concepts to 30 concepts. The video frames related to these 30 concepts in SIN11 are used to help learn the intermediate representation. We also enlist the performance variance of SCR as it also leverages the SIN dataset to obtain a concepts-based representation for MED. The first three events, *i.e.*, *Attempting a board trick*, *Feeding an animal* and *Landing a fish* are used as showcases. Table 2 displays the corresponding results. It can be seen that the performance of SAIR does not vary much when using only 30 concepts for intermediate representation. However, the performance of SCR drops drastically. For example, SCR outperforms SAIR for the event *Attempting a board trick* when using 65 concepts but



(a) *Performance Comparison in terms of MinNDC*



(b) *Performance Comparison in terms of AP*

Figure 1. Performance comparison between using 500 negative examples and using 1000 negative examples. Note that LOWER MinNDC/HIGHER AP indicates BETTER performance.

SAIR beats SCR when using 30 concepts. Thus, our method SAIR is more robust to the selection of concepts-based videos compared to SCR.

4.5 Using More Negative Examples

We further conduct an experiment to evaluate whether negative examples contribute much to the detection accuracy by increasing the number of negative examples to 1000. Figure 1 shows the performance comparison between using 500 negative examples and 1000 negative examples. It can be seen that using 1000 negative examples is clearly better than merely using 500 negative examples, which indicates that negative examples do help improve the detection accuracy. Since negative examples are quite easy to obtain in the real world, it is reasonable and beneficial to leverage such free resources for boosted detection accuracy.

4.6 Parameter Sensitivity

In our experiments we have tuned the regularization parameter α in (11). Thus, we conduct an experiment to study how the parameter α in (11) affects the detection performance. Similarly, we use *Attempting a board trick*, *Feeding an animal*, *Landing a fish* in this experiment. Figure 2 demonstrates the performance variation *w.r.t* α. For these three events, the best results are obtained when α is small.

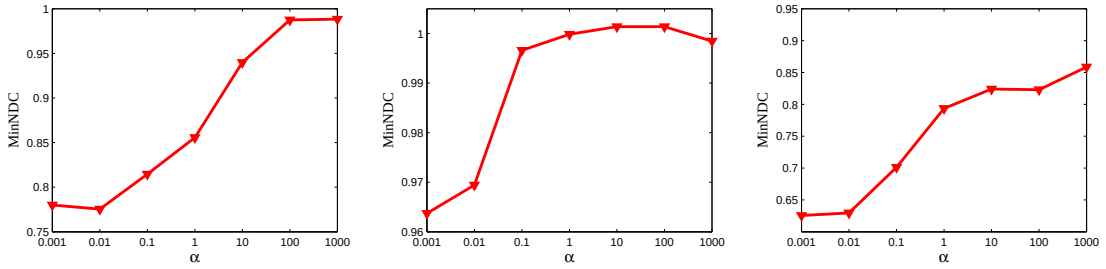
4.7 Convergence

In the previous section, we have proved that the objective function in (11) converges through the proposed algorithm. For practical applications it is interesting how fast our algorithm converges. In our convergence experiment we fix α at 1.

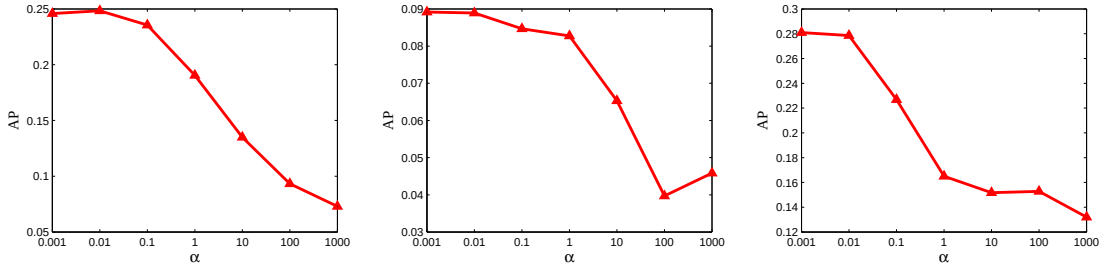
Figure 3 shows the convergence curve of our optimization algorithm. It can be seen that our algorithm converges within 10 iterations, which is efficient.

4.8 Nonlinear SAIR vs Linear SAIR

We have mentioned before that usually nonlinear classifiers obtain better performance than linear classifiers for event detection. For better performance, we have extended our algorithm SAIR to a nonlinear classifier. To understand the performance improvement from linear method to nonlinear method, we use the linear SAIR for MED. The comparison



(a) *Attempting a board trick* (b) *Feeding an animal* (c) *Landing a fish*



(d) *Attempting a board trick* (e) *Feeding an animal* (f) *Landing a fish*

Figure 2. Performance variation *w.r.t.* different values of α. Note that LOWER MinNDC/HIGHER AP indicates BETTER performance.

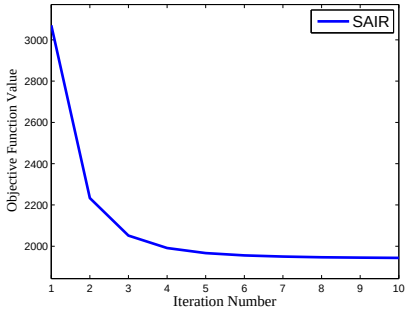


Figure 3. Convergence curve of the proposed algorithm.

between the two approaches is displayed in Figure 4. It can be seen that nonlinear SAIR has remarkable advantage over linear SAIR in terms of MinNDC and AP. The result demonstrates that it is beneficial to implement our method as a nonlinear classifier for MED.

5 Conclusion

Multimedia event detection is important for video indexing and retrieval. We have proposed a new learning framework for multimedia event detection by leveraging the classifier-specific intermediate representation from low-level features. The intermediate representation of videos is automatically optimized together with the classifier. As a result, the intermediate representation is able to better reveal the video semantics and at the same time is preferable for the classifier learning. Specifically, we have used external videos in the learning process, which provide extra informative cues. The joint learning of the intermediate representation and the classifier results in a respectable framework for multimedia event detection. To validate its efficacy, we conducted several experiments using real-world video archives. The results showed that our method consistently yields competitive or better accuracy than other methods.

6 Acknowledgments

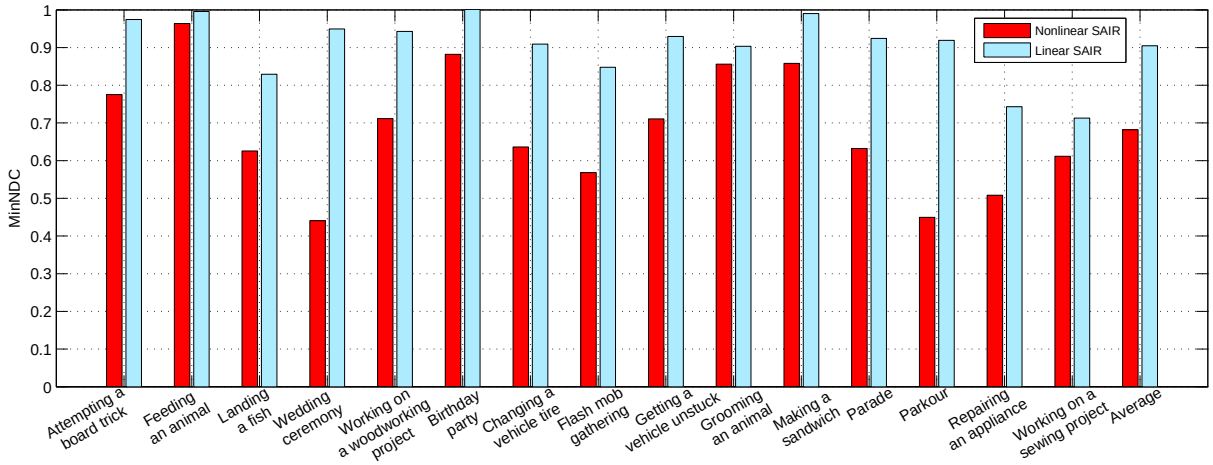
This paper was partially supported by the European Commission under contract FP7-248984 GLOCAL, the EIT ICT Labs EventMAP project and the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center contract number D11PC20068. The U.S. Government is authorized 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, or the U.S. Government. We also want to thank the Carnegie Mellon University Parallel Data Lab for the computing resources.

References

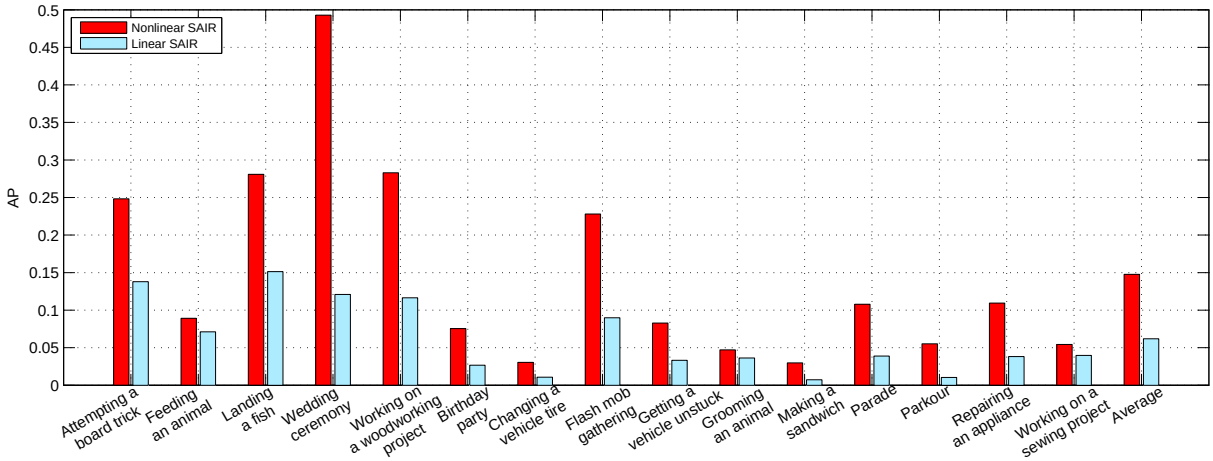
[1] A. G. Hauptmann, R. Yan, W.-H. Lin, M. G. Christel, and H. D. Wactlar. Can high-level concepts fill the semantic gap in video retrieval? a case study with broadcast news. *IEEE Transactions on Multimedia*, 9(5):958–966, 2007.

[2] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-based image retrieval at the end of the early years. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(12):1349-1380, 2000.

[3] Z. Ma, Y. Yang, Y. Cai, N. Sebe, and A. G. Hauptmann. Knowledge Adaptation for Ad Hoc Multimedia Event Detection with Few Exemplars. In *ACM MM*, 2012.



(a) *Performance Comparison in terms of MinNDC*



(b) *Performance Comparison in terms of AP*

Figure 4. Performance comparison between using nonlinear SAIR and using linear SAIR. Note that LOWER MinNDC/HIGHER AP indicates BETTER performance.

[4] C. Snoek, M. Worring, J. van Gemert, J.-M. Geusebroek, and A. W. M. Smeulders. The challenge problem for automated detection of 101 semantic concepts in multimedia. In *ACM Multimedia*, pages 421–430, 2006.

[5] M.-L. Shyu, Z. Xie, M. Chen, and S.-C. Chen. Video semantic event/concept detection using a subspace-based multimedia data mining framework. *IEEE Transactions on Multimedia*, 10(2):252–259, 2008.

[6] V. S. Tseng, J.-H. Su, J.-H. Huang, and C.-J. Chen. Integrated mining of visual features, speech features, and frequent patterns for semantic video annotation. *IEEE Transactions on Multimedia*, 10(2):260–267, 2008.

[7] Trec video retrieval evaluation. National Institute of Standards and Technology. *http://www-nlpir.nist.gov/projects/trecvid/*.

[8] M. R. Naphade and J. R. Smith. On the detection of semantic concepts at trecvid. In *ACM Multimedia*, pages 660–667, 2004.

[9] A. F. Smeaton, P. Over, and W. Kraaij. Evaluation campaigns and TRECVid. In *ACM MIR*, pages 321–330, 2006.

[10] J. Yang, R. Yan, and A. G. Hauptmann. Cross-domain video concept detection using adaptive SVMs. In *ACM Multimedia*, pages 188–197, 2007.

[11] K.-H. Liu, M.-F. Weng, C.-Y. Tseng, Y.-Y. Chuang, and M.-S. Chen. Association and temporal rule mining for post-filtering of semantic concept detection in video. *IEEE Transactions on Multimedia*, 10(2):240–251, 2008.

[12] D. A. Sadlier and N. E. O'Connor. Event detection in field sports video using audio-visual features and a support vector machine. *IEEE Trans. Circuits Syst. Video Techn.*, 15(10):1225–1233, 2005.

[13] F. Wang, Y.-G. Jiang, and C.-W. Ngo. Video event detection using motion relativity and visual relatedness. In *ACM Multimedia*, pages 239–248, 2008.

[14] A. Adam, E. Rivlin, I. Shimshoni, and D. Reinitz. Robust real-time unusual event detection using multiple fixed-location monitors. *IEEE Trans. Pattern Anal. Mach. Intell.*, 30(3):555–560, 2008.

[15] W. Hu, N. Xie, L. Li, X. Zeng, and S. J. Maybank. A survey on visual content-based video indexing and retrieval. *IEEE Transactions on Systems, Man, and Cybernetics, Part C*, 41(6):797–819, 2011.

[16] M. S. Lew, N. Sebe, C. Djeraba, and R. Jain. Content-based multimedia information retrieval: State of the art and challenges. *TOMCCAP*, 2(1):1–19, 2006.

[17] Y. Yang, F. Nie, D. Xu, J. Luo, Y. Zhuang, and Y. Pan. A multimedia retrieval framework based on semi-supervised ranking and relevance feedback. *IEEE Trans. Pattern Anal. Mach. Intell.*, 34(4):723–742, 2012.

[18] Z.-J. Zha, M. Wang, Y.-T. Zheng, Y. Yang, R. Hong, and T.-S. Chua. Interactive video indexing with statistical active learning. *IEEE Transactions on Multimedia*, 14(1):17–27, 2012.

[19] Z. Ma, Y. Yang, A. G. Hauptmann, and N. Sebe. Classifier-specific Intermediate Representation for Multimedia Tasks. In *ICMR*, 2012.

[20] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[21] I. Laptev and T. Lindeberg. Space-time interest points. In *ICCV*, pages 432–439, 2003.

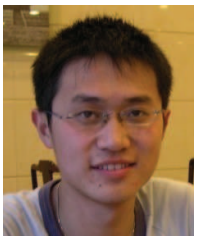
[22] M.-Y. Chen and A. Hauptmann. Mosift: Recognizing human actions in surveillance videos. In *Technical Report CMU-CS-09-161, Carnegie Mellon University*, 2009.

[23] Z.-Z. Lan, L. Bao, S.-I. Yu, W. Liu, and A. G. Hauptmann. Double fusion for multimedia event detection. In *MMM*, pages 173–185, 2012.

[24] Y. Yang, Y. Zhuang, F. Wu, and Y. Pan. Harmonizing hierarchical manifolds for multimedia document semantics understanding and cross-media retrieval. *IEEE Transactions on Multimedia*, 10(3):437–446, 2008.
[25] S. N. P. Vitaladevuni, P. Natarajan, R. Prasad, and P. Natarajan. Efficient orthogonal matching pursuit using sparse random projections for scene and video classification. In *ICCV*, pages 2312–2319, 2011.
[26] M. R. Naphade, J. R. Smith, J. Tesic, S.-F. Chang, W. H. Hsu, L. S. Kennedy, A. G. Hauptmann, and J. Curtis. Large-scale concept ontology for multimedia. *IEEE MultiMedia*, 13(3):86–91, 2006.
[27] A. G. Hauptmann, M. G. Christel, and R. Yan. Video retrieval based on semantic concepts. *Proceedings of the IEEE*, 96(4):602–622, 2008.
[28] S. Ji and J. Ye. Linear dimensionality reduction for multi-label classification. In *IJCAI*, pages 1077–1082, 2009.
[29] E. Kokiopoulou, and Y. Saad. Orthogonal Neighborhood Preserving Projections: A Projection-Based Dimensionality Reduction Technique. *IEEE Trans. Pattern Anal. Mach. Intell.*, 29(12):2143-2156, 2007.
[30] B. Schölkopf, A. J. Smola, and K.-R. Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10(5):1299–1319, 1998.
[31] C. Zhang, F. Nie, and S. Xiang. A general kernelization framework for learning algorithms based on kernel pca. *Neurocomputing*, 73(4-6):959–967, 2010.
[32] M. J. Saberian, H. Masnadi-Shirazi, and N. Vasconcelos. Taylorboost: First and second-order boosting algorithms with explicit margin control. In *CVPR*, pages 2929–2934, 2011.
[33] K. Fukunaga. *Introduction to statistical pattern recognition (2nd ed.)*. Academic Press Professional, San Diego, USA, 1990.
[34] D. Ding, F. Metze, S. Rawat, P. F. Schulam, S. Burger, E. Younessian, L. Bao, M. G. Christel and A. G. Hauptmann. Beyond audio and video retrieval: towards multimedia summarization. In *ICMR*, 2012.
[35] K. E. A. van de Sande, T. Gevers, and C. G. M. Snoek. Evaluating color descriptors for object and scene recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 32(9):1582–1596, 2010.
[36] http://www.nist.gov/itl/iad/mig/upload/med11-evalplan-v03-20110801a.pdf.

Zhigang Ma received the B.S. and M.S. both from Zhejiang University, Hangzhou, China in 2004 and 2006 respectively, and is currently working toward the PhD degree from the University of Trento, Trento, Italy. He had been an intern at the School of Computer Science, Carnegie Mellon University, Pittsburgh, PA from September, 2011 to March, 2012.

His research interests include machine learning and its application to computer vision and multimedia analysis.



Yi Yang received the Ph.D degree in Computer Science from Zhejiang University, Hangzhou, China, in 2010.

He had been a postdoctoral research fellow at the University of Queensland from 2010 to May, 2011. After that, he joined Carnegie Mellon University. He is now a Postdoctoral Research Fellow at the School of Computer Science, Carnegie Mellon University, Pittsburgh, PA. His research interests include machine learning and its applications to multimedia content analysis and computer vision, e.g. multimedia indexing and retrieval, image annotation, video semantics understanding, etc.

Nicu Sebe (M'01-SM'11) received the Ph.D. in computer science from Leiden University, Leiden, The Netherlands, in 2001.

Currently, he is with the Department of Information Engineering and Computer Science, University of Trento, Italy, where he is leading the research in the areas of multimedia information retrieval and human-computer interaction in computer vision applications. He was involved in the organization of the major conferences and workshops addressing the computer vision and human-centered aspects of multimedia information retrieval, among which as a General Co-Chair of the IEEE Automatic Face and Gesture Recognition Conference, FG 2008, ACM International Conference on Image and Video Retrieval (CIVR) 2007 and 2010, and WIAMIS 2009 and as one of the initiators and a Program Co-Chair of the Human-Centered Multimedia track of the ACM Multimedia 2007 conference. He is the general chair of ACM Multimedia 2013 and was a program chair of ACM Multimedia 2011. He is a senior member of IEEE and of ACM.

Kai Zheng is currently a Research Fellow with the Data Engineering and Pattern Recognition Research Division at the University of Queensland, Australia. He received his Bachelor and Master degrees in Computer Science from Tongji University in 2006 and Fudan University in 2009, and PhD degree in Computer Science from The University of Queensland in 2012. He had been a visiting scholar in MSRA from August to November in 2010 and 2011 respectively. His research interests include efficient query processing indexing in spatio-temporal databases, uncertain data management, and human mobility computation.

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.