Knowledge Adaptation for Ad Hoc Multimedia Event Detection with Few Exemplars *

Zhigang Ma
DISI, University of Trento, Italy
ma@disi.unitn.it

Yi Yang
SCS, Carnegie Mellon University, USA
yiyang@cs.cmu.edu

Yang Cai
SCS, Carnegie Mellon University, USA
caiyang@cs.cmu.edu

Nicu Sebe
DISI, University of Trento, Italy
sebe@disi.unitn.it

Alexander G. Hauptmann
SCS, Carnegie Mellon University, USA
alex@cs.cmu.edu

ABSTRACT

Multimedia event detection (MED) has a significant impact on many applications. Though video concept annotation has received much research effort, video event detection remains largely unaddressed. Current research mainly focuses on sports and news event detection or abnormality detection in surveillance videos. Our research on this topic is capable of detecting more complicated and generic events. Moreover, the curse of reality, *i.e.*, precisely labeled multimedia content is scarce, necessitates the study on how to attain respectable detection performance using only limited positive examples. Research addressing these two aforementioned issues is still in its infancy. In light of this, we explore Ad Hoc MED, which aims to detect complicated and generic events by using few positive examples. To the best of our knowledge, our work makes the first attempt on this topic. As the information from these few positive examples is limited, we propose to infer knowledge from other multimedia resources to facilitate event detection. Experiments are performed on real-world multimedia archives consisting of several challenging events. The results show that our approach outperforms several other detection algorithms. Most notably, our algorithm outperforms SVM by 43% and 14% comparatively in Average Precision when using Gaussian and χ^2 kernel respectively.

Categories and Subject Descriptors

H.3.1 [**Information Search and Retrieval**]: Content Analysis and Indexing; I.2.10 [**Vision and Scene Understanding**]: Video analysis

General Terms

Algorithms, Experimentation, Performance

Keywords

Multimedia Event Detection (MED), Ad Hoc MED, Knowledge Adaptation, Structural Adaptive Regression (SAR)

*Area chair: Gang Hua

MM'12, October 29–November 2, 2012, Nara, Japan.

1. INTRODUCTION

With ever expanding multimedia collections, multimedia content analysis is becoming a fundamental research issue for many applications such as indexing and retrieval, *etc.* Multimedia content analysis aims to learn the semantics of multimedia data. To do so, it has to bridge the semantic gap between the low-level features and the high-level semantic content description [11, 33]. Different approaches have been proposed to bridge the semantic gap in the literature, either at concept level or event level.

We first highlight the difference between a concept and an event. A "concept" means an abstract or general idea inferred from specific instances, *e.g. fish*, *sky*. In multimedia research, a major thrust for multimedia content analysis is to learn the semantic concepts of the multimedia data and to use these concepts for multimedia indexing and retrieval. Multimedia concept analysis has been widely studied for images and videos [17, 25, 24, 15]. However, as shared personal video collections, news videos and documentary videos have explosively proliferated these years, video event analysis gradually is attracting more research interest. An "event" refers to an observable occurrence that interests users, *e.g. making a cake*, *landing a fish*. Compared with concept analysis, where concepts are usually describable by a single shot, event understanding is a more challenging task due to its dynamic attribute and semantic richness. For example, the event *making a cake* consists of a combination of several concepts such as *cake*, *people*, *kitchen* together with the action *making* within a longer video sequence.

Annotation and detection are two different topics of both concept and event analysis. Multimedia annotation, also known as recognition, aims to associate multimedia data with one or multiple semantic labels (tags). For an image/video to be annotated, it is assigned to a specific concept or event that is already known. Many approaches have been proposed to improve the annotation accuracy for both images and videos [17, 30, 26]. The detection task, however, is different from annotation in that it aims to detect the existence of concepts or events through pre-trained detectors. Compared with annotation, detection is more challenging. In detection, there may only be a few positive examples while the negative examples come from an infinite semantic space. We have no clue about all the concepts or events these negative examples include. This provides limited information for obtaining a robust detector.

The TREC Video Retrieval Evaluation (TRECVID) community [3] has notably contributed to the research of video concept or event detection by providing a common testbed for evaluating different detection approaches [20]. In the field of multimedia, many other works have also focused on *concept detection*, e.g., [25, 32, 14]. However, the research on video *event detec-*

tion is still in its infancy. Most existing research on event detection is limited to the events in sports [22, 31, 24] and news video archives [29], or those with repetitive patterns like *running* [28] or unusual events in surveillance videos [4]. In 2010, the TRECVID community launched the task of "Event detection in Internet multimedia (MED)" which aims to encourage new technologies for detecting more complicated events, *e.g.*, *landing a fish*. The definition of the MED task from National Institute of Standards and Technology (NIST) is: 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. Though few systems have been designed for the MED task [6, 7], they only focus on predefined events. In 2012, NIST proposed an even more challenging problem of MED. The problem is how to attain respectable detection accuracy when the system is not optimized for a limited set of known events and with very few positive examples (*a.k.a.* Ad Hoc MED[1]) since precisely labeled training data are difficult to obtain in the real world. In this paper, we focus on designing a novel algorithm for Ad Hoc MED which deals with the limited number of positive training examples. To the best of our knowledge, this work is the first research on Ad Hoc MED.

Ad Hoc MED faces two major challenges, *i.e.*, complicated events and few positive examples. SVM has been shown to be the most effective tool for predefined MED [6, 7, 13]. However, it is not suitable for Ad Hoc MED when there are only a few positive examples. Since there are some available video archives with annotated concept labels, we can leverage them to facilitate Ad Hoc MED. The difficulty is that those concepts are different from the event to be detected. Hence, a method is in demand to bridge the gap between the concepts and the event, thus being able to utilize the concepts-based videos. Inspired by [32, 12, 8], we propose to adapt the knowledge from concept level to assist in Ad Hoc MED. Specifically, we use the available video corpora with annotated concepts as our auxiliary resource and Ad Hoc MED is performed on the target videos. The concepts are relevant to the event to be detected.

The main contributions of our work are as follows:

(1) We perform the first exploration of Ad Hoc MED research by proposing a novel approach built atop knowledge adaptation.

(2) Unlike many knowledge adaptation methods, our approach does not require that auxiliary videos have the same events as the target videos. We exploit videos with several semantic *concepts* to facilitate the Ad Hoc *Event* Detection on the target videos; the event differs from the concepts and video collections are different from each other.

(3) Compared to other detectors, leveraging knowledge from an auxiliary video archive enables us to obtain improved detection rates in the target video archive with only few positive examples.

2. RELATED WORK

In this section, we briefly review the related works on video semantic analysis and knowledge adaptation.

2.1 Video Semantic Analysis

In the past, although multimedia event analysis has been less focused, video concept annotation has been widely studied. For instance, in [26] Tseng *et al.* have proposed using integrated mining of visual features, speech features, and frequent semantic patterns of videos for annotation. Besides, Snoek *et al.* have studied the challenging problem of automatically indexing 101 semantic concepts [25]. Their work also provides the research community with a manually annotated lexicon containing 101 semantic concepts. In [14], a general post-filtering framework using association and temporal analysis has been proposed for concept classification.

Event detection is a more challenging problem, which has not been sufficiently studied. Most of the existing research efforts are limited to the detection of sports events, news events, unusual surveillance events or those with repetitive patterns. For example, Xu *et al.* propose using web-casting text and broadcast video to detect events from the live sports game [31]. In [29], a model based on a multi-resolution, multi-source and multi-modal bootstrapping framework has been developed for events detection in news videos. Adam *et al.* present an algorithm using multiple local monitors which collect low-level statistics to detect certain types of unusual events in surveillance videos [4]. Sports events, news events and unusual events are usually predefined so that we can identify some event-specific rules or templates to facilitate detection of the particular event. For example, to detect the event *goal* in sports videos, we can utilize people's cheers as a strong evidence. However, in Ad Hoc MED events are more generic and we do not know what the events are before conducting the detection task. Thus, the aforementioned methods may not work well for Ad Hoc MED. Wang *et al.* have proposed a new motion feature by using motion relativity and visual relatedness for event detection [28]. Their approach primarily applies to events that have repetitive motion attributes and are usually describable by a single shot, *e.g. walking* and *dancing*. In contrast, Ad Hoc MED focuses on events that have varying motion attributes within a longer video sequence. For instance, *making a cake* includes different motions such as getting the flour, adding water and baking within a longer video sequence. Moreover, there are few positive examples for training. Thus, the approach in [28] is unsuitable for Ad Hoc MED.

More recently, some researchers began to study the predefined MED [6, 7, 13]. Compared to sports events, news events, unusual surveillance events or those with repetitive patterns, the events in predefined MED are more complicated and difficult to detect. In predefined MED, SVM is widely used and shows good performance. As opposed to predefined MED, we focus on an even more difficult problem Ad Hoc MED, which detects generic events. However, SVM is not suitable for Ad Hoc MED due to the "curse of reality", *i.e.*, few precisely labeled positive examples are provided. Therefore, effective algorithms are in demand for Ad Hoc MED to promote video semantic analysis to a more mature level.

2.2 Knowledge Adaptation for Multimedia Analysis

Knowledge adaptation, also known as transfer learning, aims to propagate the knowledge from an auxiliary domain to a target domain [32, 12, 8, 10]. A number of algorithms have been proposed but most of them require that the auxiliary domain and the target domain have the same classes. However, Ad Hoc MED deals with very complicated events which are not predefined and come from unlimited semantic space. Hence, most existing methods are not applicable. For example, Yang *et al.* have proposed to use Adaptive SVMs for cross-domain video concept detection [32]. The method obtained encouraging results but has some shortcomings. The proposed approach requires that the auxiliary videos and the target videos have the same video concepts. However, in Ad Hoc MED the event to be detected is unknown before we perform the detection task. Collecting many auxiliary videos with the same event description as the target videos is impossible. Jiang *et al.* [12] have used the image context of Flickr to select concept detectors. These pre-selected detectors are then refined by the semantic context transfer from the target domain. In this way, more precise concept detectors are obtained for video search. The proposed method

[1] http://www.nist.gov/itl/iad/mig/med12.cfm

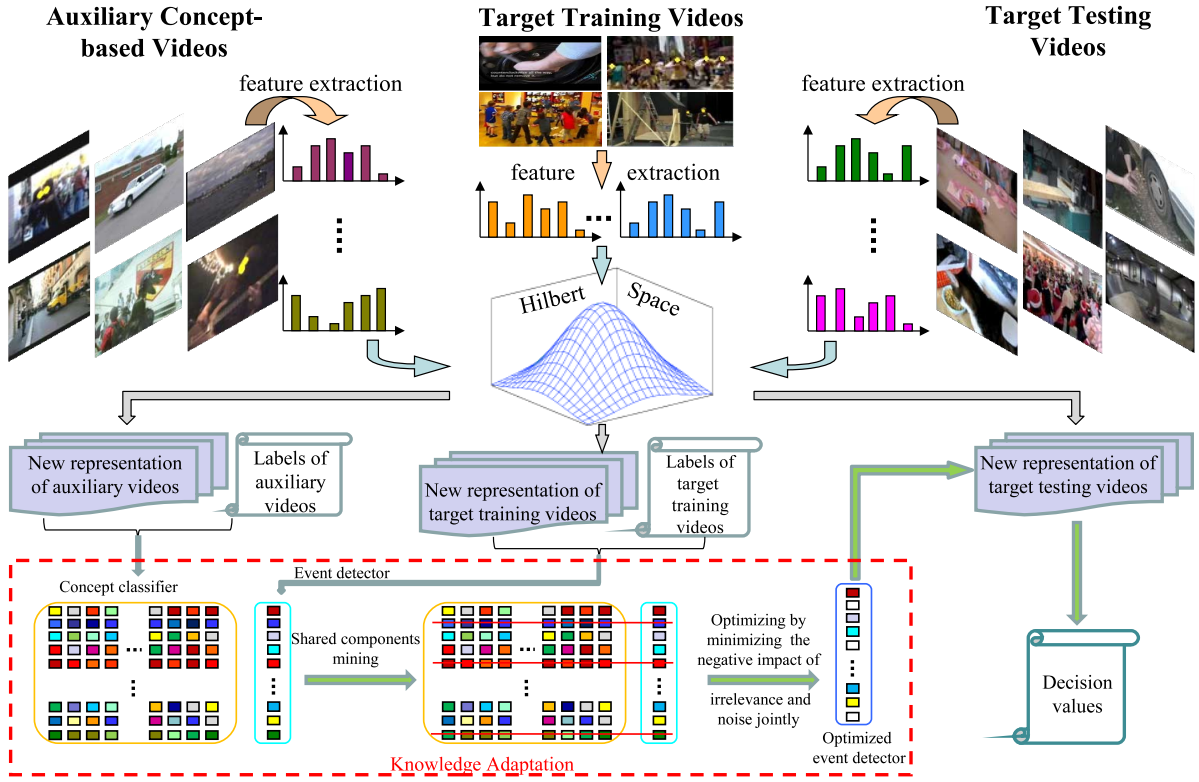


Figure 1: The illustration of our framework. We first map the low-level features of the auxiliary and target videos into a Hilbert Space. The video concept classifier and the video event detector presumably have common components which contain irrelevance and noise. We propose to remove such negative information by optimizing the concept classifier and the event detector jointly.

is interesting but the selected concept detectors cannot be handily used for event detection without other sophisticated algorithms. Besides, as in Ad Hoc MED we only have very few positive examples, using these examples to refine the concept detectors is not reliable. Another algorithm proposed by Duan *et al.* [8] realizes event recognition of consumer videos by leveraging web videos. Their method does not require that the auxiliary domain and the target domain have the same events. However, the approach is very time consuming so it is not suitable for real-world applications.

To progress beyond these aforementioned works, we propose a new knowledge adaptation method for Ad Hoc MED. It explores the shared knowledge between the concepts-based auxiliary videos and the event-based target videos. The shared knowledge can be used to refine the detector of the target videos for Ad Hoc MED. Different from the state of the art, our algorithm does not require that the auxiliary videos have the same event as the target videos. It is also computationally efficient.

3. VIDEO REPRESENTATION

Now we illustrate our framework for Ad Hoc MED. Figure 1 shows the framework of our approach. The video archive where the Ad Hoc MED is to be conducted is our target domain. The low-level features of both auxiliary and target videos are mapped into a Hilbert Space where the shared knowledge between them is to be explored. The video concept classifier and the video event detector presumably have common components which contain irrelevance and noise. We propose to remove such components by optimizing the concept classifier and the event detector jointly, thereby resulting in a more discriminative event detector when we have very few positive examples.

We first introduce the video representation in this section. It has been demonstrated that using multiple features for multimedia representation always yields better performance [34, 13]. We use the SIFT feature [16] and CSIFT feature [27] for video representation. Considering computational efficiency, we use a shot boundary detection algorithm to extract key frames. The shot boundary detection algorithm works as follows: First, it calculates the color histogram of every 5 frames; Second, it subtracts the histogram with that of the previous frame; Third, the frame will be a shot boundary if the subtracted value is larger than an empirically set threshold. Once we get the shot, the frame in the middle of the shot is used to represent that shot. Then we use the Harris-Laplace interest point detector to detect interest points. The SIFT/CSIFT descriptor is subsequently used and we obtain a 4096 dimension Bag-of-Words feature, for which we sum over all the interest points in a video, which is then normalized. The SIFT and CSIFT features are further combined so we use an 8192 dimension feature to represent each video.

Suppose there are n_t training videos for event detector training. The training videos consist of both positive and negative examples. Denote $X_t = [x_t^1, x_t^2, ..., x_t^{n_t}] \in \mathbb{R}^{d \times n_t}$ as the BoW feature of these training videos. d is the feature dimension ($d = 8192$ in this paper). As it has been reported that kernelization is an effective way to deal with BoW features for video analysis [6], after extracting BoW features from the videos, we leverage kernel tricks to transform x_t^i to $\tilde{x}_t^i$. Specifically, we perform full rank principal component analysis in another Hilbert Space H, which is related to the input space by a nonlinear map $\Phi : \mathbb{R}^d \to H$. The covariance ma-

trix in H is given by$C_H = \frac{1}{n_t}\sum_{i=1}^{n_t}\Phi(x_t^i)\Phi(x_t^i)^T$ [23], where x_t^i is the i^{th} training MED video sequence. We aim to find eigenvalues $\lambda \geq 0$ and eigenvectors V satisfying $\lambda V = CV$. Although H could have an arbitrarily large, possibly infinite dimensionality, the inner product of any two data $\Phi(x_i)$ and $\Phi(x_j)$ can be explicitly expressed by a kernel matrix K, *i.e.*, $K_{ij} = \big(\Phi(x_i)\cdot\Phi(x_j)\big)$. It turns out that we need to solve the eigenvalue problem $n_t\lambda\alpha = K\alpha$, where $\alpha = [\alpha_1, ..., \alpha_{n_t}]^T$ are coefficients such that $V = \sum_{i=1}^{n_t}\alpha_i\Phi(x_t^i)$ [23]. Let $\alpha^1, ..., \alpha^r$ be the normalized eigenvectors corresponding to all the non-zero eigenvalues $0 < \lambda_1 \leq ... \leq \lambda_r$. For a testing video sequence x_{te} and an auxiliary video sequence x_a, the projection into the eigenvectors $V^p (1 \leq p \leq r)$ in H can be computed according to $\sum_{i=1}^{n_t}\alpha_i^p\left(\Phi(x_t^i)\cdot\Phi(x_{te})\right)$ and $\sum_{i=1}^{n_t}\alpha_i^p\left(\Phi(x_t^i)\cdot\Phi(x_a)\right)$.

4. CONCEPTS ADAPTATION ASSISTED EVENT DETECTION

Next, we explain how the knowledge adaptation is accomplished for Ad Hoc MED. Our approach is grounded on two components: one is the knowledge from the available target training examples and the other one is the knowledge propagated from the auxiliary concepts-based videos.

We first demonstrate how to exploit the knowledge from the target training examples. Denote the target training videos in H by $\tilde{X}_t = [\tilde{x}_t^1, \tilde{x}_t^2, ..., \tilde{x}_t^{n_t}] \in \mathbb{R}^{d_h \times n_t}$. $y_t = [y_t^1, y_t^2, ..., y_t^{n_t}]^T \in \{0,1\}^{n_t \times 1}$ are the labels for the target training videos. $y_t^i = 1$ if the i^{th} video x_t^i is a positive example whereas $y_t^i = 0$ otherwise. The low-level representations and high-level semantics of videos can be associated by a decision function f which, for an input video sequence x, predicts an output y. In this paper, we define f_t as:

$$f_t(\tilde{X}_t^i) = \tilde{X}_t^T W_t + 1_t b_t, \tag{1}$$

where $W_t \in \mathbb{R}^{d_h \times 1}$ is an event detector which correlates $\tilde{X}_t$ with their labels y_t. $b_t \in \mathbb{R}^1$ is a bias term and $1_t \in \mathbb{R}^{n_t \times 1}$ denotes a column vector with all ones. f_t is decided by minimizing the following objective based on the training examples X_t and their labels y_t:

$$\min_{f_t} loss\left(f_t(X_t), y_t\right) + \mu\Omega(f_t). \tag{2}$$

$loss(\cdot)$ is a loss function and $\alpha\Omega(f_t)$ is the regularization function on f_t with μ as its parameter. Different loss functions such as the hinge loss and the least square loss can be used. In this paper, we use the $\ell_{2,1}$-norm based loss function because it is robust to outliers [18]. Thus, Eq. (2) is reformulated as:

$$\min_{W_t, b_t} \left\| \tilde{X}_t^T W_t + 1_t b_t - y_t \right\|_{2,1} + \mu\Omega(W_t). \tag{3}$$

Now we show how to adapt the knowledge from auxiliary videos which are associated with different concepts to assist in Ad Hoc MED. Denote the auxiliary videos in H by $\tilde{X}_a = [\tilde{x}_a^1, \tilde{x}_a^2, ..., \tilde{x}_a^{n_a}] \in \mathbb{R}^{d_h \times n_a}$. $Y_a = [y_a^1, y_a^2, ..., y_a^{n_a}]^T \in \{0,1\}^{n_a \times c_a}$ is their label matrix where c_a indicates that there are c_a different concepts. Y_a^{ij} denotes the j^{th} datum of y_a^i and $Y_a^{ij} = 1$ if x_a^i belongs to the j^{th} concept, while $Y_a^{ij} = 0$ otherwise. The fundamental step is to mine the correlation between the low-level representations and high-level semantics of the auxiliary concepts-based videos. Similarly to Eq. (3), we realize that by the following objective:

$$\min_{W_a, b_a} \left\| \tilde{X}_a^T W_a + 1_a b_a - Y_a \right\|_{2,1} + \gamma\Omega(W_a) \tag{4}$$

where a concept classifier $W_a \in \mathbb{R}^{d_h \times c_a}$ is used to correlate $\tilde{X}_a$ with their labels Y_a. $b_a \in \mathbb{R}^{1 \times c_a}$ is a bias term and $1_a \in \mathbb{R}^{n_a \times 1}$ is a column vector with all ones.

Next, we illustrate how to adapt knowledge from the auxiliary concepts-based videos for a more discriminating event detector. The concepts-based videos presumably share some common components with the target event-based videos. For instance, the concepts *fish*, *water*, *people* are relevant with the event *landing a fish*. It is reasonable to leverage such relevance to improve the target event detection when we have only few positive examples. As the classifier W_a and the detector W_t correlate the low-level representations with the high-level concepts and event for each domain respectively, we explore the shared knowledge between the two to optimize the learning of W_t. The video representation in H is potentially noisy. In Ad Hoc MED, only few training examples are provided. The limited information is usually not sufficient to effectively deal with the underlying noise. On the other hand, the concepts of $\tilde{X}_a$ and the event of $\tilde{X}_t$ are related and grounded on similar low-level representations. The irrelevant or noisy components in W_a and W_t should be similar, which can be uncovered by learning W_a and W_t jointly. Thus, we exploit the concept classifier W_a to help remove the noise in W_t for a more discriminative event detector.

Denote $W_a = [w_a^1, ..., w_a^{d^h}]^T$, $W_t = [w_t^1, ..., w_t^{d^h}]^T$. Then we combine them and define a joint analyzer $W = [w^1, ..., w^{d^h}]$ where w^i is the horizontal concatenation of w_a^i and w_t^i, *i.e.*, $w^i = [w_a^i, w_t^i]$. In this sense, w_i reflects the joint information from the auxiliary videos and the target training videos. Through proper optimization of w_i, we can remove the shared irrelevant or noisy components. Previous work has shown that sparse models are useful for feature selection by eliminating redundancy and noise [5, 19, 18]. The sparse models are used to make some of the feature coefficients shrink to zeros to achieve feature selection. These works, though focusing on different problems, provide us with the inspiration that the "shrinking to zero" idea can be applied to uncover the common structures shared by the concept classifier and the event detector. In this way, we can similarly remove the shared irrelevance and noise, thus obtaining a more discriminative event detector. Specifically, we propose to minimize $\|W\|_{2,p} = \left(\sum_{i=1}^{d_h}\left(\sum_{j=1}^{c_a+1}\left|W_{ij}\right|\right)^{\frac{p}{2}}\right)^{\frac{1}{p}}$ to achieve that goal. $\|\cdot\|_{2,p}$ denotes the $\ell_{2,p}$-norm ($0 < p < 2$). By minimizing $\|W\|_{2,p}$, we can reduce the negative impact of the irrelevant or noisy w_i's. p is used to control the degree of shared structures. The lower p is, the more correlated are the concept classifier and the event detector. Consequently, we can obtain an optimal event detector W_t.

To this end, we propose the following objective function to adapt the knowledge from auxiliary videos for Ad Hoc MED:

$$\begin{aligned}\min_{W_a, W_t, b_a, b_t} & \left\|\tilde{X}_a^T W_a + 1_a b_a - Y_a\right\|_{2,1} + \left\|\tilde{X}_t^T W_t + 1_t b_t - y_t\right\|_{2,1} \\ & + \alpha\left(\sum_{i=1}^{d_h}\left(\sum_{j=1}^{c_a+1}\left|W_{ij}\right|\right)^{\frac{p}{2}}\right)^{\frac{1}{p}} + \beta(\|W_a\|_F^2 + \|W_t\|_F^2)\end{aligned} \tag{5}$$

where $\beta(\|W_a\|_F^2 + \|W_t\|_F^2)$ is added to control the capacity of the classifier and the detector.

Once W_t is obtained, we apply it to the testing videos in H for event detection. Our method builds upon the knowledge adaptation from concepts-based videos to event-based videos by leveraging the shared structures between them. We therefore name it Structural Adaptive Regression (SAR).

5. OPTIMIZING THE EVENT DETECTOR

In this section, we present our solution for obtaining the target event detector. Our problem in Eq. (5) involves the $\ell_{2,1}$-norm and the $\ell_{2,p}$-norm which are both non-smooth and cannot be solved in

a closed form. We propose to solve it as follows. For the detailed solution, please see Appendix.

Denote $\tilde{X}_a^T W_a - Y_a = [u^1, ..., u^{n_a}]^T$, $\tilde{X}_t^T W_t - y_t = [v^1, ..., v^{n_t}]^T$. Next, we define three diagonal matrices D_a, D_t and D with their diagonal elements $D_a^{ii} = \frac{1}{2\|u^i\|_2}$, $D_t^{ii} = \frac{1}{2\|v^i\|_2}$, $D^{ii} = \frac{1}{\frac{2}{p}\|w^i\|_2^{2-p}}$ respectively. In this way, we can get the concept classifier W_a as:

$$W_a = (\tilde{X}_a H_a D_a H_a \tilde{X}_a^T + \alpha D + \beta I_d)^{-1} \tilde{X}_a H_a D_a H_a Y_a \quad (6)$$

$H_a = I_a - \frac{1}{n_a} 1_a 1_a^T$. $I_a \in \mathbb{R}^{n_a \times n_a}$ and $I_d \in \mathbb{R}^{d_h \times d_h}$ are two identity matrices. The event detector W_t is obtained as:

$$W_t = (\tilde{X}_t H_t D_t H_t \tilde{X}_t^T + \alpha D + \beta I_d)^{-1} \tilde{X}_t H_t D_t H_t y_t. \quad (7)$$

where $H_t = I_t - \frac{1}{n_t} 1_t 1_t^T$ and $I_t \in \mathbb{R}^{n_t \times n_t}$ is an identity matrix.

Next, we propose Algorithm 1 to solve the objective function in Eq. (5). It can be proved that the objective function in Eq. (5) is convex and the objective function value monotonically decreases in each iteration until convergence using Algorithm 1. Due to the space limit, we omit the proof. For training, the computational complexity of Algorithm 1 is $O(d_h^3)$. Note that $d_h \leq n_t$ because usually there are few training examples in Ad Hoc MED. Thus, the training process is not very computationally expensive. During testing, computing kernels between the testing data and the training data is the most expensive process. Suppose there are n_{te} testing videos, we need to compute $n_t n_{te}$ kernels. Each datum is d_h dimensional so the complexity is $O(d_h n_t n_{te})$.

Algorithm 1: Optimizing the event detector.

Input:

The auxiliary data $\tilde{X}_a \in \mathbb{R}^{d_h \times n_a}, Y_a \in \mathbb{R}^{n_a \times c_a}$;

The target training data $\tilde{X}_t \in \mathbb{R}^{d_h \times n_t}, y_t \in \mathbb{R}^{n_t \times 1}$;

Parameters α and β.

Output:

Optimized $W_t \in \mathbb{R}^{d_h \times 1}$ and $b_t \in \mathbb{R}^1$.

1: Set $t = 0$, initialize $W_a \in \mathbb{R}^{d_h \times c_a}$ and $W_t \in \mathbb{R}^{d_h \times 1}$ randomly;

2: **repeat**

Compute $\tilde{X}_a^T W_a - Y_a = [u^1, ..., u^{n_a}]^T$, $\tilde{X}_t^T W_t - y_t = [v^1, ..., v^{n_t}]^T$, and $W = [w^1, ..., w^d]^T$;

Compute the diagonal matrix D_a^t, D_t^t and D^t according to $D_a^{ii} = \frac{1}{2\|u^i\|_2}$, $D_t^{ii} = \frac{1}{2\|v^i\|_2}$, and $D^{ii} = \frac{1}{\frac{2}{p}\|w^i\|_2^{2-p}}$ respectively;

Update W_a^{t+1} as:
$W_a^{t+1} = (\tilde{X}_a H_a D_a \tilde{X}_a^T + \alpha D + \beta I_d)^{-1} \tilde{X}_a H_a D_a H_a Y_a^T$;

Update b_a^{t+1} as: $b_a^{t+1} = \frac{1}{n_a} 1_a^T Y_a - \frac{1}{n_a} 1_a^T \tilde{X}_a^T W_a^{t+1}$;

Update W_t^{t+1} as:
$W_t^{t+1} = (\tilde{X}_t H_t D_t \tilde{X}_t^T + \alpha D + \beta I_d)^{-1} \tilde{X}_t H_t D_t H_t y_t^T$;

Update b_t^{t+1} as: $b_t^{t+1} = \frac{1}{n_t} 1_t^T y_t - \frac{1}{n_t} 1_t^T \tilde{X}_t^T W_t^{t+1}$;

$t = t + 1$.

until *Convergence*;

3: Return W_t and b_t.

6. EXPERIMENTS

In this section, we present the experiments which aim to evaluate the performance of our Structural Adaptive Regression (SAR) for Ad Hoc MED.

6.1 Datasets

NIST has provided so far the largest video corpora for MED. Our experiments on Ad Hoc MED are conducted on the TRECVID MED 2010 (MED10) and TRECVID MED 2011 (MED11) development set. MED10[2] includes 3 events defined by NIST, which are *Making a cake*, *Batting a run*, and *Assembling a shelter*. MED11[3] includes 15 events, *i.e.*, *Attempting a board trick*, *Feeding an animal*, *Landing a fish*, *Wedding ceremony*, *Working on a woodworking project*, *Birthday party*, *Changing a vehicle tire*, *Flash mob gathering*, *Getting a vehicle unstuck*, *Grooming an animal*, *Making a sandwich*, *Parade*, *Parkour*, *Repairing an appliance* and *Working on a sewing project*. The two datasets are combined together (MED10-11 for short) in our experiments so we have a dataset of 9822 video clips consisting of 361,623 key frames.

We use the development set from TRECVID 2011 semantic indexing task (SIN11) as the auxiliary videos. The SIN11 covers 346 concepts but some of them have few positive examples. Additionally, "events" usually refer to "semantically meaningful human activities, taking place within a selected environment and containing a number of necessary objects [9]." Hence, we removed the concepts with few positive examples and selected 65 concepts that are related to human, environment and objects. We thus use a subset with 2529 video frames.

6.2 Setup

The videos are represented by the SIFT and CSIFT features. 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.

According to the MED task definition from NIST, each event is detected independently. Therefore, there are 18 individual detection tasks. NIST has defined that the number of positive training examples is 10 for Ad Hoc MED [2]. However, there is no standard training and testing set partition provided by NIST. Hence, we randomly split the MED10-11 dataset into two subsets, one as the training set and the other one as the testing set. We follow the Ad Hoc MED definition by NIST and randomly select 10 positive examples for each event. Another 300 negative examples are selected and combined with the positive examples as the training data. The remaining 9512 videos are our testing data. The experiments are independently repeated 5 times with randomly selected positive and negative examples. The average results are reported.

We use three evaluation metrics. The first one, Minimum NDC (MinNDC), is officially used by NIST in TRECVID MED 2011 evaluation [1]. Lower MinNDC indicates better detection performance. The second one is the Probability of Miss-Detection based on the Detection Threshold 12.5. This evaluation metric is used by NIST in TRECVID MED 2012 [2] to evaluate MED performance. We denote it Pmd@TER=12.5 for short. Likewise, lower Pmd@TER=12.5 indicates better performance. For more details about the above two evaluation metrics, please see the TRECVID 2011 and 2012 evaluation plans [1, 2]. The third one is Average Precision (AP). Higher AP indicates better performance.

6.3 Ad Hoc MED Results

In this section, we show the Ad Hoc MED results. As SVM is the most widely used and robust event detector for MED [13, 6, 11, 28], we first compare our method SAR with SVM. We use two kernels, *i.e.*, Gaussian kernel and χ^2 kernel for both SAR and SVM. To be clear, we use G-SAR, G-SVM, χ^2-SAR and χ^2-SVM to refer

[2]http://nist.gov/itl/iad/mig/med10.cfm

[3]http://www.nist.gov/itl/iad/mig/med11.cfm

Table 1: Average detection accuracy of SAR and SVM using Gaussian kernel. LOWER MinNDC and Pmd@TER=12.5 indicate BETTER performance. HIGHER AP indicates BETTER performance. Better results are highlighted in bold.

Evaluation Metric	G-SVM	G-SAR
MinNDC	0.954	**0.910**
Pmd@TER=12.5	0.751	**0.674**
AP	0.072	**0.103**

Table 2: Average detection accuracy of SAR and SVM using χ^2 kernel. LOWER MinNDC and Pmd@TER=12.5 indicate BETTER performance. HIGHER AP indicates BETTER performance. Better results are highlighted in bold.

Evaluation Metric	χ^2-SVM	χ^2-SAR
MinNDC	0.904	**0.881**
Pmd@TER=12.5	0.665	**0.626**
AP	0.115	**0.131**

to different implementations. For SAR, we tune the two parameters α and β both from $\{10^{-3}, 10^{-2}, ..., 10^2, 10^3\}$, and the parameter p from $\{0.5, 1, 1.5\}$. For SVM, we use LIBSVM and tune the parameters C and γ similarly from $\{10^{-3}, 10^{-2}, ..., 10^2, 10^3\}$. We report the best results for each algorithm.

Figure 2 shows the comparison between the four approaches. Note that LOWER MinNDC and Pmd@TER=12.5 indicate BETTER performance; HIGHER AP indicates BETTER performance. We observe that: 1) In general, χ^2-SAR and χ^2-SVM are better than G-SAR and G-SVM respectively, demonstrating that χ^2 kernel is better than Gaussian kernel for video analysis. 2) χ^2-SAR proposed in this paper is consistently the most competitive algorithm. Specifically, χ^2-SAR attains the best detection performance for 14, 14 and 13 events in MinNDC, Pmd@TER=12.5 and AP respectively; χ^2-SAR gets the advantageous performance for the remaining events. 3) For average performance over the 18 events, G-SAR, χ^2-SAR outperform G-SVM, χ^2-SVM respectively using all the three evaluation metrics. The detailed numbers of the average results are listed in Table 1 and Table 2. 4) We also notice that the performance of G-SAR does not change significantly compared with χ^2-SAR. However, for most events, the performance of G-SVM drops drastically compared with χ^2-SVM, which indicates that SVM is sensitive to kernel change. In fact, the advantage of G-SAR over G-SVM is remarkably large for many events. For few events, *e.g.*, *Working on a woodworking project* and *Grooming an animal*, χ^2-SAR is worse than χ^2-SVM. Nonetheless, G-SAR outperforms G-SVM dramatically.

Next, we add two state of the art detectors for comparison:

(1) TaylorBoost [21]: a state of the art algorithm extended from AdaBoost.

(2) Adaptive Multiple Kernel Learning (A-MKL) [8]: a recent knowledge adaptation algorithm built upon SVM.

For these algorithms, we use the code shared by the authors. A-MKL is a knowledge adaptation based algorithm, which similarly utilizes the SIN11 dataset as auxiliary data.

Since χ^2-SAR and χ^2-SVM show better performance in the last experiment, we only list their results to compare with the other three approaches. Note that we use SAR and SVM for short here.

Figure 3 shows the detection results of different approaches. Note that lower MinNDC and Pmd@TER=12.5 indicate better performance; higher AP indicates better performance. We can see that our method SAR is still the most competitive algorithm. Specifically, SAR is the best for 14, 13 and 10 events in MinNDC, Pmd@TER=12.5 and AP respectively. For the remaining events, SAR obtains the advantageous performance. For average performance, SAR is the best algorithm using all the three evaluation metrics. We also observe that A-MKL generally attains the second best performance. However, we would point out that A-MKL is not suitable for large scale multimedia analysis due to its low computational efficiency. The calculation of 80 different kernel matrices consumes up to 30GB memory and the detection for one event is much more computationally expensive than our algorithm.

6.4 Using Fewer Concepts

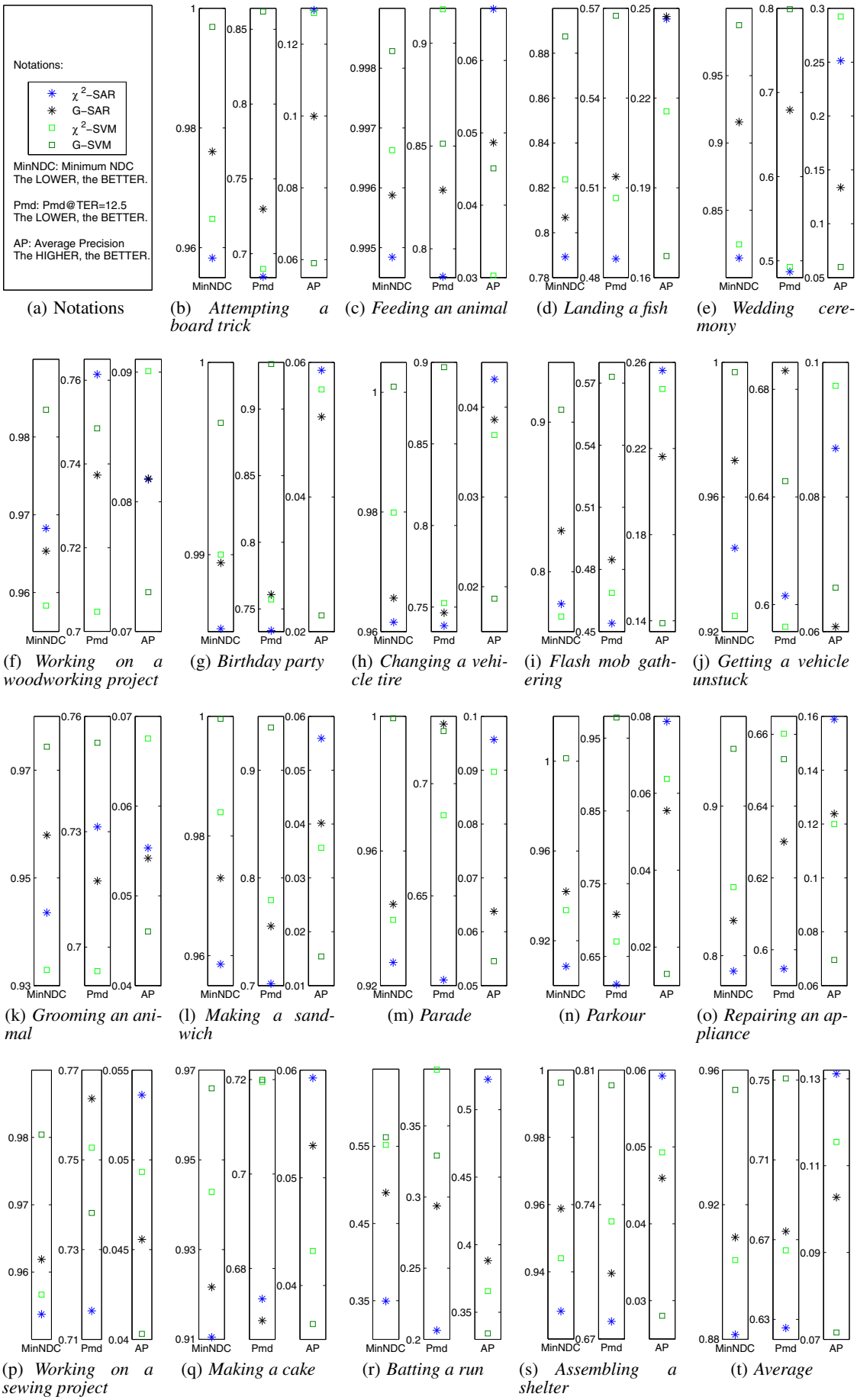
To study whether the number of concepts affects the Ad Hoc MED performance, we conduct an experiment by choosing fewer concepts out of the 65 concepts from the auxiliary videos. We manually selected 30 concepts which are supposed to be most related to generic events. The videos related to these 30 concepts are used as auxiliary data. Figure 4 displays the corresponding results. We only show the results in Average Precision due to the space limit. It can be seen that the performance does not vary much when using only 30 auxiliary concepts. This observation indicates that it is not very critical to decide how many concepts should be selected as auxiliary knowledge with our method.

6.5 Do Negative Examples Help?

We further conduct an experiment to evaluate whether negative examples contribute much to the detection accuracy by reducing the number of negative examples to 100. Figure 5 shows the performance comparison between using 100 negative examples and using 300 negative examples. Similarly, Average Precision is chosen as the evaluation metric. It can be seen that using 300 negative examples is clearly better than merely using 100 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 cheap resources for boosted detection accuracy.

7. CONCLUSION

In this paper, we have introduced the first research exploration of Ad Hoc MED. This is an important research issue as it focuses on more generic, complicated and meaningful events that reflect our daily activities. In addition, the situation we are faced in the real world requires that only few positive examples are used. To achieve good performance, we have proposed to borrow strength from available concepts-based videos for Ad Hoc MED. In our joint optimization framework, we first mine the shared irrelevance and noise between the auxiliary videos and the target videos. Then a sophisticated method is exerted to alleviate the negative impact of the irrelevance and noise to obtain a more robust event detector. We also proposed an efficient iterative algorithm to solve our objective function. Extensive experiments using real-world multimedia archives were conducted with results showing that our method generally outperforms all compared state of the art detection algorithms. This promising performance indicates that it is beneficial to leverage auxiliary knowledge for Ad Hoc MED when we do not have sufficient positive examples. However, knowledge adaptation is based on the assumption that there are shared structures between the source and the target. If the two have very different structures, we may get negative transfer which degrades the detection performance on the target. Therefore, it would be interesting and important to study how to judge the structural commonness to better utilize knowledge adaption in the future.



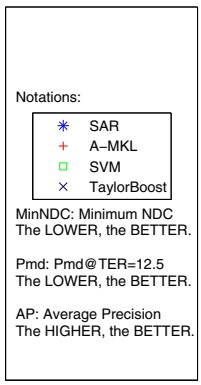
(a) Notations (b) *Attempting a board trick* (c) *Feeding an animal* (d) *Landing a fish* (e) *Wedding ceremony*

(f) *Working on a woodworking project* (g) *Birthday party* (h) *Changing a vehicle tire* (i) *Flash mob gathering* (j) *Getting a vehicle unstuck*

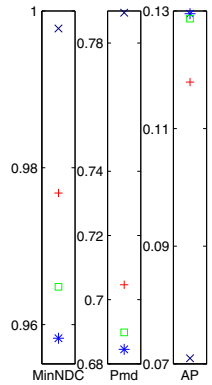
(k) *Grooming an animal* (l) *Making a sandwich* (m) *Parade* (n) *Parkour* (o) *Repairing an appliance*

(p) *Working on a sewing project* (q) *Making a cake* (r) *Batting a run* (s) *Assembling a shelter* (t) *Average*

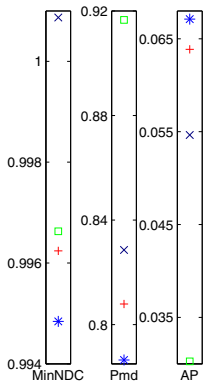
Figure 2: Performance Comparison between SAR and SVM using different kernels. Note that LOWER MinNDC and Pmd@TER=12.5 indicate BETTER performance; HIGHER AP indicates BETTER performance.



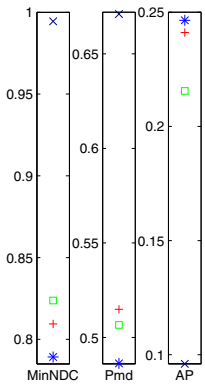
(a) Notations



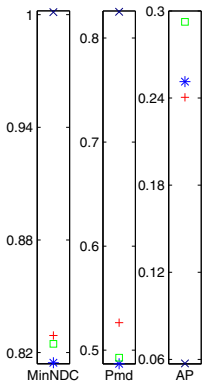
(b) *Attempting a board trick*



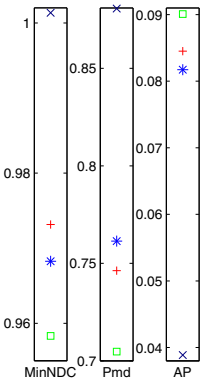
(c) *Feeding an animal*



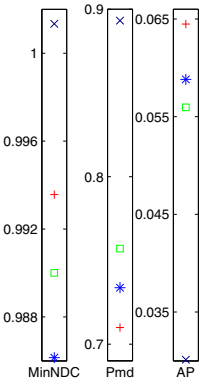
(d) *Landing a fish*



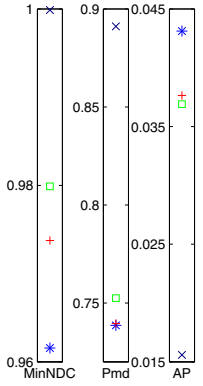
(e) *Wedding ceremony*



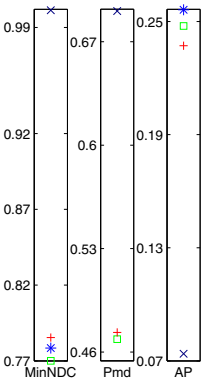
(f) *Working on a woodworking project*



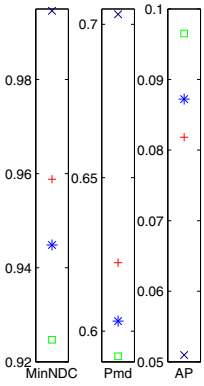
(g) *Birthday party*



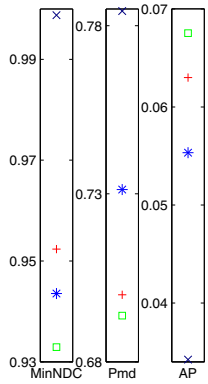
(h) *Changing a vehicle tire*



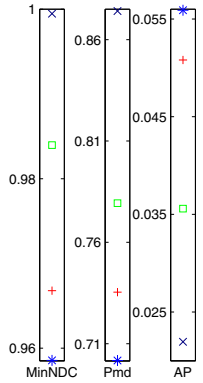
(i) *Flash mob gathering*



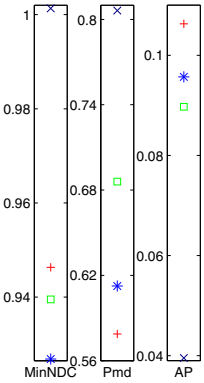
(j) *Getting a vehicle unstuck*



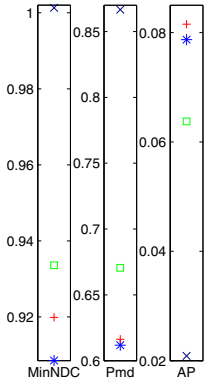
(k) *Grooming an animal*



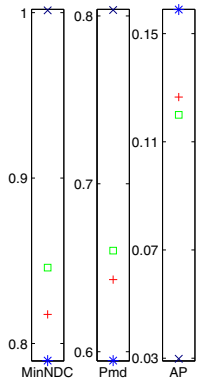
(l) *Making a sandwich*



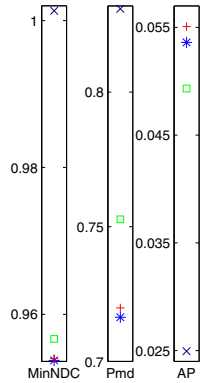
(m) *Parade*



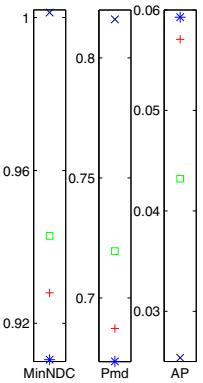
(n) *Parkour*



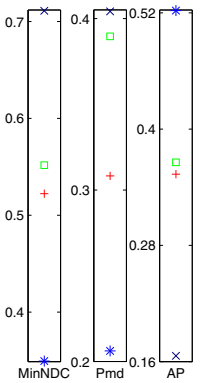
(o) *Repairing an appliance*



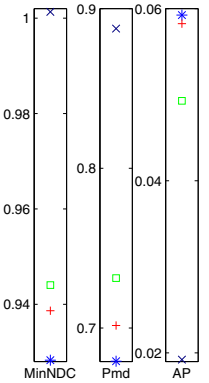
(p) *Working on a sewing project*



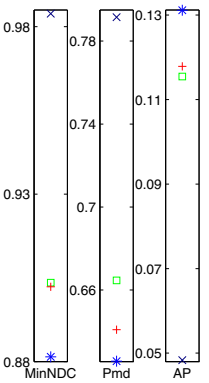
(q) *Making a cake*



(r) *Batting a run*



(s) *Assembling a shelter*



(t) *Average*

Figure 3: Performance Comparison on Ad Hoc MED. Note that LOWER MinNDC and Pmd@TER=12.5 indicate BETTER performance; HIGHER AP indicates BETTER performance.

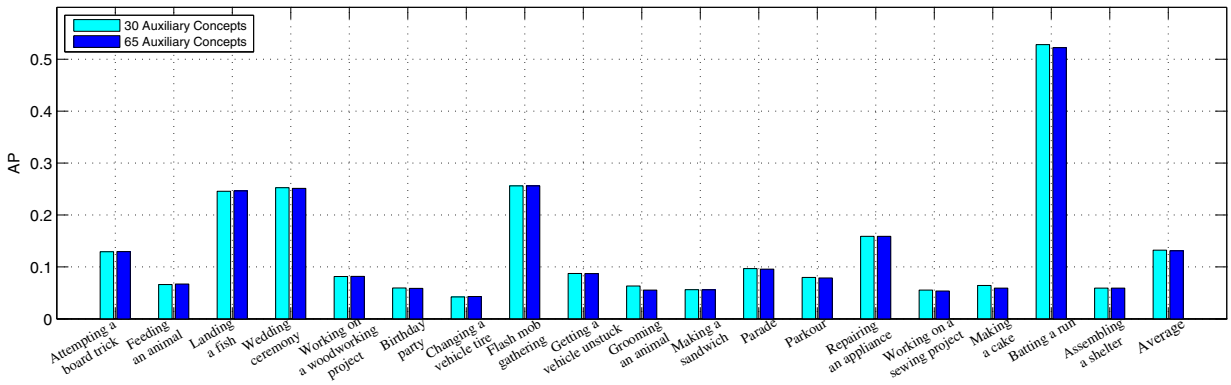


Figure 4: Performance comparison between using 30 auxiliary concepts and using 65 auxiliary concepts.

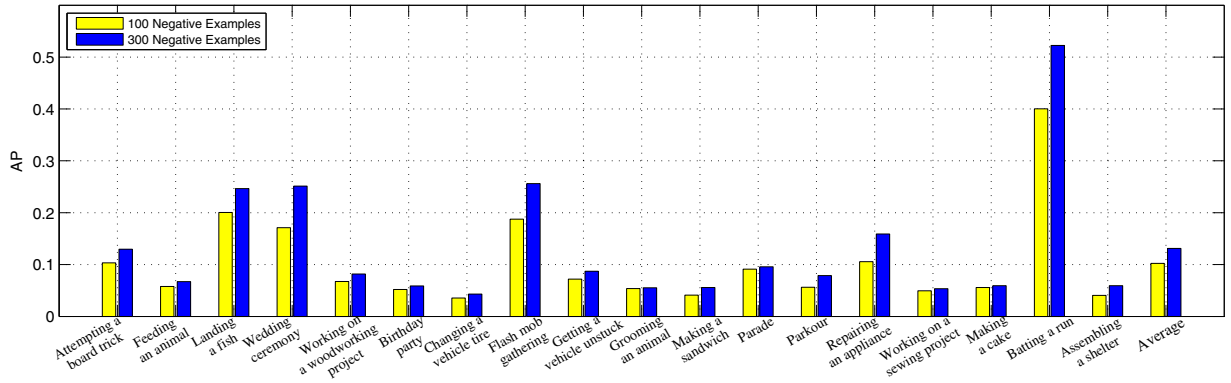


Figure 5: Performance comparison between using 100 negative examples and using 300 negative examples.

8. ACKNOWLEDGMENTS

This paper was partially supported by the Glocal FP7 IP project, the S-PATTERNS FIRB 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.

9. REFERENCES

[1] http://www.nist.gov/itl/iad/mig/upload/med11-evalplan-v03-20110801a.pdf.

[2] http://www.nist.gov/itl/iad/mig/upload/med12-evalplan-v01.pdf.

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

[4] 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.

[5] A. Argyriou, T. Evgeniou, and M. Pontil. Multi-task feature learning. In *NIPS*, pages 41–48, 2006.

[6] L. Bao, L. Zhang, S.-I. Yu, Z. zhong Lan, L. Jiang, A. Overwijk, Q. Jin, S. Takahashi, B. Langner, Y. Li, M. Garbus, S. Burger, F. Metze, and A. Hauptmann. Informedia @ TRECVID2011. In *NIST TRECVID Workshop*, 2011.

[7] L. Cao, S.-F. Chang, N. Codella, C. Cotton, D. Ellis, L. Gong, M. Hill, G. Hua, J. Kender, M. Merler, Y. Mu, A. Natsev, and J. R. Smith. IBM Research and Columbia University TRECVID-2011 Multimedia Event Detection (MED) System. In *NIST TRECVID Workshop*, 2011.

[8] L. Duan, D. Xu, I. W.-H. Tsang, and J. Luo. Visual event recognition in videos by learning from web data. In *CVPR*, 2010.

[9] L. Fei-Fei and L.-J. Li. What, Where and Who? Telling the Story of an Image by Activity Classification, Scene Recognition and Object Categorization. *Studies in Computational Intelligence- Computer Vision*, pages 157–171, 2010.

[10] Y. Han, F. Wu, Y. Zhuang, and X. He. Multi-label transfer learning with sparse representation. *IEEE Trans. Circuits Syst. Video Techn.*, 20(8):1110–1121, 2010.

[11] 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.

[12] Y.-G. Jiang, C.-W. Ngo, and S.-F. Chang. Semantic context transfer across heterogeneous sources for domain adaptive video search. In *ACM Multimedia*, pages 155–164, 2009.

[13] 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.

[14] 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.

[15] Y. Liu, F. Wu, Y. Zhuang, and J. Xiao. Active post-refined multimodality video semantic concept detection with tensor representation. In *ACM Multimedia*, pages 91–100, 2008.

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

[17] J. Luo, J. Yu, D. Joshi, and W. Hao. Event recognition: viewing the world with a third eye. In *ACM Multimedia*, pages 1071–1080, 2008.

[18] Z. Ma, F. Nie, Y. Yang, J. Uijlings, and N. Sebe. Web image annotation via subspace-sparsity collaborated feature selection. *IEEE Transactions on Multimedia*, 2012.

[19] Z. Ma, Y. Yang, F. Nie, J. R. R. Uijlings, and N. Sebe. Exploiting the entire feature space with sparsity for automatic image annotation. In *ACM Multimedia*, pages 283–292, 2011.

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

[21] M. J. Saberian, H. Masnadi-Shirazi, and N. Vasconcelos. Taylorboost: First and second-order boosting algorithms with explicit margin control. In *CVPR*, 2011.

[22] 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.

[23] B. Schölkopf, A. J. Smola, and K.-R. Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10(5):1299–1319, 1998.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] 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.

[29] G. Wang, T.-S. Chua, and M. Zhao. Exploring knowledge of sub-domain in a multi-resolution bootstrapping framework for concept detection in news video. In *ACM Multimedia*, pages 249–258, 2008.

[30] M. Wang, X.-S. Hua, X. Yuan, Y. Song, and L.-R. Dai. Optimizing multi-graph learning: towards a unified video annotation scheme. In *ACM Multimedia*, pages 862–871, 2007.

[31] C. Xu, J. Wang, K. Wan, Y. Li, and L. Duan. Live sports event detection based on broadcast video and web-casting text. In *ACM Multimedia*, pages 221–230, 2006.

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

[33] 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.

[34] 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.

APPENDIX

The objective in Eq. (5) is equivalent to:

$$\begin{aligned}&\min_{W_a,W_t,b_a,b_t} Tr\left((\tilde{X}_a^T W_a + 1_a b_a - Y_a)^T D_a (\tilde{X}_a^T W_a + 1_a b_a - Y_a)\right)\\&+Tr\left((\tilde{X}_t^T W_t + 1_t b_t - y_t)^T D_t (\tilde{X}_t^T W_t + 1_t b_t - y_t)\right)\\&+\alpha Tr\left(W^T D W\right) + \beta(\|W_a\|_F^2 + \|W_t\|_F^2)\end{aligned} \tag{8}$$

where $Tr(\cdot)$ denotes the trace operator. By setting the derivative of Eq. (8) *w.r.t.* b_a to zero, we get:

$$\begin{aligned}&21_a^T D_a \tilde{X}_a^T W_a + 21_a^T D_a 1_a b_a - 21_a^T D_a Y_a = 0\\&\Rightarrow b_a = \frac{1}{n_a} 1_a^T Y_a - \frac{1}{n_a} 1_a^T \tilde{X}_a^T W_a.\end{aligned} \tag{9}$$

Similarly, we obtain b_t as:

$$b_t = \frac{1}{n_t} 1_t^T y_t - \frac{1}{n_t} 1_t^T \tilde{X}_t^T W_t. \tag{10}$$

Substituting Eq. (9) and Eq. (10) into Eq. (8), it becomes:

$$\begin{aligned}&\min_{W_a,W_t} Tr\left([\tilde{X}_a^T W_a + 1_a(\frac{1}{n_a}1_a^T Y_a - \frac{1}{n_a}1_a^T \tilde{X}_a^T W_a) - Y_a]^T D_a\right.\\&\left.[\tilde{X}_a^T W_a + 1_a(\frac{1}{n_a}1_a^T Y_a - \frac{1}{n_a}1_a^T \tilde{X}_a^T W_a) - Y_a]\right)\\&+Tr\left([\tilde{X}_t^T W_t + 1_t(\frac{1}{n_t}1_t^T y_t - \frac{1}{n_t}1_t^T \tilde{X}_t^T W_t) - y_t]^T D_t\right.\\&\left.[\tilde{X}_t^T W_t + 1_t(\frac{1}{n_t}1_t^T y_t - \frac{1}{n_t}1_t^T \tilde{X}_t^T W_t) - y_t]\right)\\&+\alpha Tr\left(W^T D W\right) + \beta(\|W_a\|_F^2 + \|W_t\|_F^2)\end{aligned} \tag{11}$$

Let $H_a = I_a - \frac{1}{n_a} 1_a 1_a^T$ and $H_t = I_t - \frac{1}{n_t} 1_t 1_t^T$ where $I_a \in \mathbb{R}^{n_a \times n_a}$ and $I_t \in \mathbb{R}^{n_t \times n_t}$ are two identity matrices. We then rewrite Eq. (11) as:

$$\begin{aligned}&\min_{W_a,W_t} Tr\left((H_a \tilde{X}_a^T W_a - H_a Y_a)^T D_a (H_a \tilde{X}_a^T W_a - H_a Y_a)\right)\\&+Tr\left((H_t \tilde{X}_t^T W_t - H_t y_t)^T D_t (H_t \tilde{X}_t^T W_t - H_t y_t)\right)\\&+\alpha Tr\left(W^T D W\right) + \beta(\|W_a\|_F^2 + \|W_t\|_F^2)\end{aligned} \tag{12}$$

Setting the derivative of Eq. (12) *w.r.t.* W_a to zero, it becomes:

$$\begin{aligned}&2\tilde{X}_a H_a D_a H_a \tilde{X}_a^T W_a + 2\alpha D W_a + 2\beta W_a - 2\tilde{X}_a H_a D_a H_a Y_a = 0\\&\Rightarrow W_a = (\tilde{X}_a H_a D_a H_a \tilde{X}_a^T + \alpha D + \beta I_d)^{-1} \tilde{X}_a H_a D_a H_a Y_a\end{aligned} \tag{13}$$

where $I_d \in \mathbb{R}^{d_h \times d_h}$ is an identity matrix. In the same manner, we obtain the event detector W_t as:

$$W_t = (\tilde{X}_t H_t D_t H_t \tilde{X}_t^T + \alpha D + \beta I_d)^{-1} \tilde{X}_t H_t D_t H_t y_t. \tag{14}$$