Knowledge Adaptation with Partially Shared Features for Event Detection Using Few Exemplars

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

Abstract—Multimedia event detection (MED) is an emerging area of research. Previous work mainly focuses on simple event detection in sports and news videos, or abnormality detection in surveillance videos. In contrast, we focus on detecting more complicated and generic events that gain more users' interest, and we explore an effective solution for MED. Moreover, our solution only uses few positive examples since precisely labeled multimedia content is scarce in the real world. As the information from these few positive examples is limited, we propose using knowledge adaptation to facilitate event detection. Different from the state of the art, our algorithm is able to adapt knowledge from another source for MED even if the features of the source and the target are partially different, but overlapping. Avoiding the requirement that the two domains are consistent in feature types is desirable as data collection platforms change or augment their capabilities and we should be able to respond to this with little or no effort. We perform extensive experiments on real-world multimedia archives consisting of several challenging events. The results show that our approach outperforms several other state-of-the-art detection algorithms.

Index Terms—Multimedia Event Detection (MED), Knowledge Adaptation, Heterogenous Features, Heterogeneous Features based Structural Adaptive Regression (HF-SAR)

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 [17][41]. 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 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. 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 [24][32][31]. However, as shared personal video collections, news videos and documentary videos have explosively proliferated these years, video event analysis is gradually attracting more research interest. An "event" refers to an observable occurrence that interests users, *e.g. celebrating the New Year*. 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 *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 a datum with one or multiple semantic labels (tags). Many approaches have been proposed to improve the annotation accuracy for both images and videos [24][33]. A typical annotation approach first pre-trains a series of classifiers, one for each class, and then applies the pre-trained classifiers to predicting the class label of each testing datum. In contrast to annotation, detection identifies the occurrence of a class of interest. One main difference between annotation and detection is that in annotation each testing datum is guaranteed to be a positive sample of one of the predefined classes while the negative examples in detection are from a set of infinite classes. In other words, both the training and testing data in annotation tasks are from a fixed number of classes but the training and testing data in detection tasks can be from an infinite number of classes. We have no clue about all the concepts or events these negative examples include. This provides very limited training information for obtaining a robust detector, thus making detection a challenging problem.

The TREC Video Retrieval Evaluation (TRECVID) com-

- *Z. Ma and A. Hauptmann are with the School of Computer Science, Carnegie Mellon University.*
E-mail: kevinma, alex@cs.cmu.edu

- *Y. Yang is with the School of Information Technology and Electrical Engineering, The University of Queensland.*
E-mail: yee.i.yang@gmail.com

- *N. Sebe is with the Department of Information Engineering and Computer Science, University of Trento.*
E-mail: sebe@disi.unitn.it

munity [3] has notably contributed to the research of video concept and event detection by providing a common testbed for evaluating different detection approaches [27]. In the field of multimedia, many other works have also focused on *concept detection*, e.g., [32][40][22]. However, the research on video *event detection* is still in its infancy. Before 2011, most existing research on event detection was limited to the events in sports [29][39][31] and news video archives [38], or those with repetitive patterns like *running* [37] or unusual events in surveillance videos [4][5][6]. In 2010, the TRECVID community launched the task of "Event detection in Internet multimedia (MED)" which aims to encourage new technologies for detecting more generic and complicated events, *e.g.*, *landing a fish*. For this kind of events, there are huge intra-class variations. Besides, they can only be characterized by long video sequences, which necessitates the exploration of all the sequences for analysis. Figure 1 shows some frames from two videos of the same event *landing a fish*. At the first glance, we may consider Video 1 to be *skiing* as it contains the *concept* of "outdoor with snow" which is not a typical scene for *landing a fish*. The scene of Video 2 is more typical, in contrast, though it can also be a scene for *sailing*. The comparison of these two videos aims to demonstrate the huge intra-class variation of complex events. On the other hand, the information from only a few frames is patchy, as shown in Figure 1. Thus, the entire video is needed for analysis.

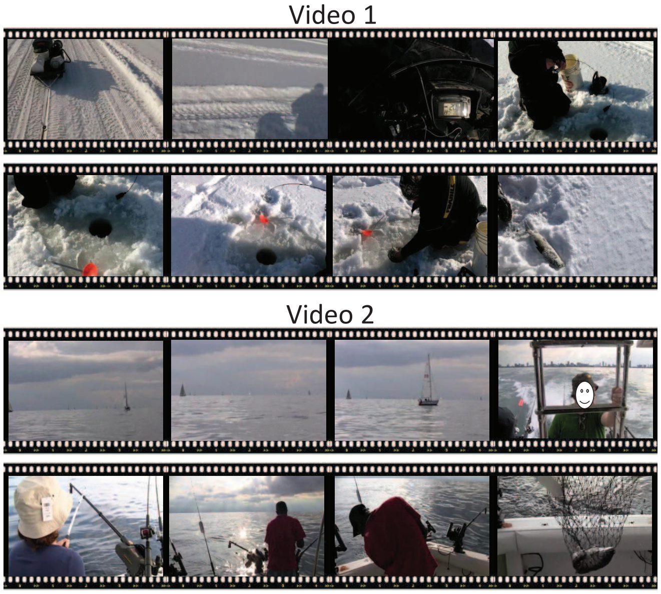


Figure 1. Some sample frames from two videos of the event *landing a fish*.

SVM has been used in few systems designed for the MED task and proved to be highly effective [10][11][19]. These systems commonly use sufficient positive examples (about 100) for reliable performance. Recently, NIST has proposed a problem of how to attain respectable detection accuracy when there are very few positive examples since precisely labeled multimedia content is scarce in the real world. In this paper, we focus on developing an effective method for MED with few exemplars. Though SVM is effective in current systems, its performance would likely be less robust when there are only a few positive examples for training. Humans often adapt knowledge obtained from previous experiences to improve learning of new tasks. Therefore, in the same manner, it is advantageous to leverage and adapt knowledge from other related domains or tasks to address the problem of an insufficient number of labeled examples. In the multimedia community, there are some available video archives with annotated concept labels, which can be leveraged to facilitate MED with few exemplars. Inspired by [40][18][12], we propose to adapt the knowledge from concept level to assist in our task. Specifically, we use the available video corpora with annotated concepts as our auxiliary resource and MED is performed on the target videos. The concepts are supposed to be relevant to the event to be detected.

Currently, most knowledge adaptation algorithms require that the features extracted from the raw data in the source domain and the target domain must be of exactly the same type. In many applications, such a requirement may be too restrictive, as data collection platforms change or augment their capabilities. In practice, the data in MED and those in the available concept-based video archives usually only have partially shared data features. For example, many video archives are key-frame based so they cannot be represented by audio features such as MFCC. These kinds of features are commonly used for MED and provide additional information for event detection. Hence, we propose to study how to effectively adapt knowledge from one domain to another when the available feature sets are partially different, but overlapping, for example if new or different features have more or better instrumentation for observations.

This paper is the extension of our previous work [13]. We summarize the main merits of this paper as follows:

- We perform an exploration of MED with few exemplars by proposing a novel approach built atop knowledge adaptation.
- 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 *event* detection on the target videos; the event differs from the concepts and the video collections are different from each other.
- Another merit is that our method is able to adapt knowledge from other sources to the target videos when only parts of the feature space are shared by the two domains. This is an intrinsic difference from most state-of-the-art knowledge adaptation algorithms.

2 Related Work

In this section, we briefly review the related work on video event detection and knowledge adaptation.

2.1 Video Event Detection

Event detection is a challenging problem that has not been yet sufficiently studied. Based on its difficulty, event detection can be roughly categorized into simple event detection, predefined MED and Ad Hoc MED.

2.1.1 Simple Event Detection

Much effort has been dedicated to the detection of sports events, news events, unusual surveillance events or those with repetitive patterns. For example, Xu *et al.* propose using webcasting text and broadcast video to detect events from live sports game [39]. In [38], a model based on a multi-resolution, multi-source and multi-modal bootstrapping framework has been developed for events detection in news videos. News videos are more constrained as they are recorded and edited by professionals. Hence, they are usually well structured and easier to analyze compared to the internet MED 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]. Wang *et al.* have proposed a new motion feature by using motion relativity and visual relatedness for event detection [37]. Their approach primarily applies to events that have repetitive motion attributes and are usually describable by a single shot, *e.g. walking* and *dancing*. The aforementioned events are usually simple, well-defined and describable by a short video sequence.

2.1.2 Multimedia Event Detection

In 2010, "Event detection in Internet multimedia (MED)" was initialized in the TRECVID competition by NIST for detecting more complicated events. Compared to the simple events mentioned above, the events in MED usually contain many people and/or objects, various human actions, multiple scenes and have significant intra-class variations. Additionally, these events take place in much longer and more complex video clips. For instance, *making a cake* includes objects such as water and bowl; can happen either in the kitchen or outdoor; is accompanied by specific motions such as getting the flour, adding water and baking within a longer video sequence. Though MED is an arduous problem, researchers have been making steady effort on it [10][11][19][20][21].

NIST introduced the predefined MED competition as follows: Each team is given the event kits about 5 months before the submission of the detection system. Hence, there is enough time for the system to be tailored particularly for a specific event. SVM is widely used and shows good performance for predefined MED. We may also use some recent state-of-the-art classifiers for MED. For example, a new family of boosting algorithms is proposed in [28] and demonstrates prominent performance on a variety of applications. In predefined MED, we can identify some event-specific rules or templates to facilitate detection of the particular event.

To address the generalizability of the MED system, NIST introduced Ad Hoc MED competition[1] in 2012. Ad Hoc MED differs from predefined MED in the sense that we should not tailor the system for a specific event. For this purpose, NIST releases the event kits to each team only about 12 days before the submission of the detection system. In this case, we know the testing events when we build the system but the short time period does not allow for special tuning for a specific event.

For both predefined MED and Ad Hoc MED, NIST has introduced an even more challenging problem, *i.e.*, using few labeled positive exemplars to build a detection system to deal with the scarcity of labeled multimedia content. Our work focuses on this problem by adapting knowledge from auxiliary concept-based data. As we do not select auxiliary concepts for a particular event, our work is different from predefined MED. Moreover, the time needed for building our system satisfies the time constraint regulated by NIST. Consequently, our work gets as close as possible to Ad Hoc MED in the intended understanding of NIST.

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 [40][18][12]. Many existing algorithms require that the features extracted from the raw data in the source domain and the target domain must be using the exact same raw sensor output. However, MED deals with very complicated events that come from an unlimited semantic space. Furthermore, the requirement of feature consistency may be too restrictive, as data collection platforms change or augment their capabilities. Hence, most existing methods are not capable of adapting knowledge for MED when we have heterogeneous feature type between the source and the target. For example, Yang *et al.* have proposed to use Adaptive SVMs for cross-domain video concept detection [40]. 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 MED the events are complicated and collecting many auxiliary videos with the same event description as the target videos within limited time is impractical. Jiang *et al.* [18] have used the image context of Flickr to select concept detectors. These pre-selected detectors are then refined by the semantic context learnt from the target domain. In this way, more precise concept detectors are obtained for video search. The proposed method is interesting but the selected concept detectors cannot be handily used for event detection without other sophisticated algorithms. Besides, as in our problem 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.* [12] 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 computation of multiple kernel is time and space consuming, especially when the feature dimension is high. Luo *et al.* have presented an object classification method by casting prior features learned from auxiliary images into their multiple kernel learning framework and obtained advantageous performance [14]. Yet this approach works in a two-step fashion, *i.e.*, training prior features using auxiliary data and then incorporating them into the following step. In contrast, our method works in a unified framework which can jointly optimize the knowledge from the auxiliary domain and the target domain. Besides those limitations mentioned above, existing knowledge adaptation algorithms mostly require that the features in the source domain and the target domain be of exactly the same type. However, in practice, this requirement

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

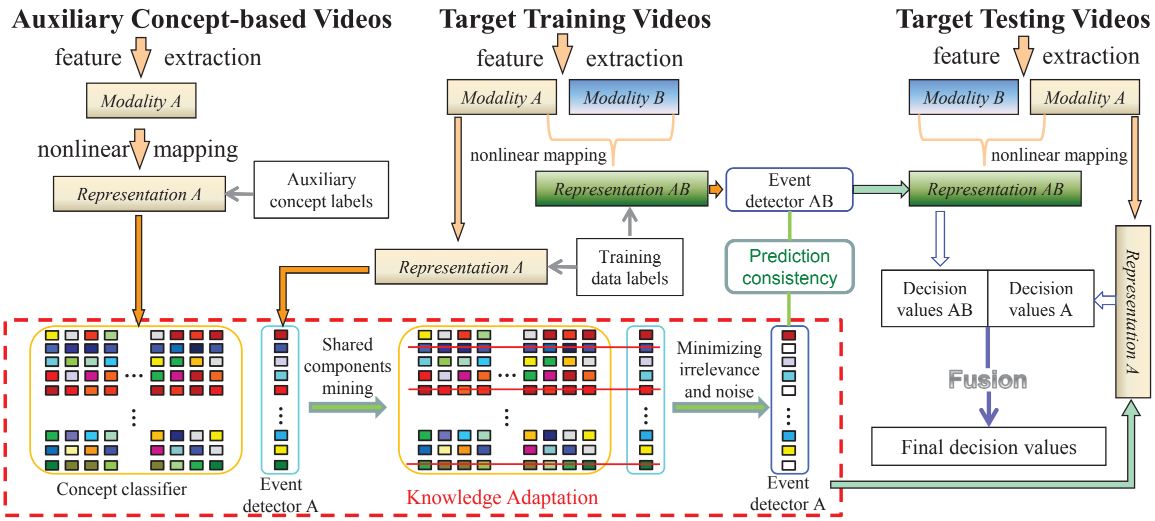


Figure 2. The illustration of our framework. We first perform nonlinear mapping for the homogeneous features of the auxiliary and target videos, *i.e.*, Modality A. The video concept classifier and the video event detector obtained from the homogeneous features presumably have common components which contain irrelevance and noise. We propose to remove such negative information by optimizing the concept classifier and the event detector jointly. Meanwhile, another event detector of MED videos is trained based on both Modality A and Modality B. Then we integrate the two event detectors for optimization, after which the decision values from both are fused for the final prediction.

may be too restrictive as MED videos can be represented by different types of features in contrast with the auxiliary video archives. Our previous work in [13] has some advantages compared to the existing knowledge adaptation algorithms such as no requirement for the same classes between the auxiliary domain and the target domain, efficiency, *etc*. But it still ignores the reality that the auxiliary domain and the target domain possibly have heterogenous feature type.

To progress beyond these aforementioned works, we propose a new knowledge adaptation method for MED with few exemplars from heterogeneous features. During the training phase, the partially shared features of the source domain and target domain will be exploited to establish a correspondence between the two domains. Meanwhile, the instrumentation obtained from the particular MED features is incorporated into our framework. The two kinds of aforementioned knowledge are then integrated to refine the detector of the target videos.

3 Framework Overview

Figure 2 illustrates our framework for MED with few exemplars. The video archive where the MED is to be conducted is our target domain. A nonlinear mapping is applied to the homogeneous features of the auxiliary and target videos, denoted by Modality A. Based on the resulting representations, the shared knowledge between them is to be explored. Specifically, we perform KPCA [30] to complete the mapping. The video concept classifier and the video event detector obtained from the homogeneous features 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 bringing discriminating knowledge for the event detector. On the other hand, we have the heterogeneous features Modality B for MED videos and they are combined with the homogeneous features as indicated in Figure 2. Another event detector of MED videos is subsequently trained based on the resulting representations from the nonlinear mapping of Modality A and Modality B. Then we integrate the two event detectors for optimization, after which the decision values from both are fused for the final prediction.

4 Concepts Adaptation assisted Event Detection

Next, we explain how we adapt knowledge for MED with few exemplars when the two domains have heterogeneous features. 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 resulting representations of the target training videos using both Modality A and Modality B after the nonlinear mapping as $\tilde{Z}_t = [\tilde{z}_t^1, \tilde{z}_t^2, ..., \tilde{z}_t^{n_t}] \in \mathbb{R}^{d_z \times n_t}$ where t stands for the target, d_z is the feature dimension and n_t is the number of the training data. $y_t = [y_t^1, y_t^2, ..., y_t^{n_t}]^T \in \{0,1\}^{n_t \times 1}$ are the labels for the target training videos. $y_t^i = 1$ if the i^{th} video is a positive example whereas $y_t^i = 0$ otherwise. To begin with, we associate the low-level representations and high-level semantics of videos by a decision function f which, for an input video sequence z, predicts an output y. In this paper, we define f_t as:

$$f_t(\tilde{Z}_t) = \tilde{Z}_t^T P_t + 1_t b_t, \tag{1}$$

where $P_t \in \mathbb{R}^{d_z \times 1}$ is an event detector which correlates $\tilde{Z}_t$ 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 $\tilde{Z}_t$ and their labels y_t:

$$\min_{f_t} loss\left(f_t(\tilde{Z}_t), y_t\right). \tag{2}$$

$loss(\cdot,\cdot)$ is a loss function. 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 [25]. Thus, Eq. (2) is reformulated as:

$$\min_{P_t,b_t} \left\|\tilde{Z}_t^T P_t + 1_t b_t - y_t\right\|_{2,1}. \tag{3}$$

Now we show how to adapt the knowledge from auxiliary videos which are associated with different concepts and are represented only by the homogeneous features, *i.e.*, Modality A to assist in MED with few exemplars. Denote the resulting representations of the auxiliary videos after the nonlinear mapping as $\tilde{X}_a = [\tilde{x}_a^1, \tilde{x}_a^2, ..., \tilde{x}_a^{n_a}] \in \mathbb{R}^{d_h \times n_a}$ where a stands for the auxiliary domain, d_h is the feature dimension and n_a is the number of the auxiliary videos. $Y_a = [y_a^1, y_a^2, ..., y_a^{n_a}]^T \in \{0,1\}^{n_a \times c_a}$ is their label matrix where c_a indicates that there are c_a different concepts. Y_a^{ij} denotes the j^{th} class of y_a^i and $Y_a^{ij} = 1$ if $\tilde{x}_a^i$ 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 function:

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

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. To begin with, we also use Modality A for the target videos in accordance with the auxiliary videos. Denote the resulting representations after the nonlinear mapping as $\tilde{X}_t = [\tilde{x}_t^1, \tilde{x}_t^2, ..., \tilde{x}_t^{n_t}] \in \mathbb{R}^{d_h \times n_t}$. We can similarly find an event detector W_t based on $\tilde{X}_t$. $W_t \in \mathbb{R}^{d_h \times 1}$ is used to correlate $\tilde{X}_t$ with their labels y_t.

Considering each domain separately, it is reasonable to assume that for classification purposes some noisy and irrelevant features will not be used, which in turn makes the corresponding rows of the projection matrix W_a or W_t identically equal to zero. Considering the two domains together, the auxiliary concept videos and the event videos can be correlated in the semantic level, *e.g.*, the concepts *fish*, *water*, *people* are basic elements of the event *landing a fish*. Previous work on multi-task learning has suggested that this kind of correlation usually results in common components in the feature level shared across related tasks [7][8][9]. In our scenario, the semantically related auxiliary videos and event videos can be treated as related tasks because the events build upon the related concepts. When we represent videos from both domains with the same type of feature such as SIFT Bag-of-Words using the same centroid, they would have some shared components. For example, assuming that the event video *landing a fish* has SIFT Bag-of-Words of *fish*, we may find similar SIFT Bag-of-Words in an image of *fish*. Hence, some shared components in the features between them need to be uncovered. Note that the event detector is actually a mapping function from features to event labels. Intuitively, not all the Bag-of-words are related to semantic labels. Given certain Bag-of-Words, if they are irrelevant to all the concepts, it is very likely that these Bag-of-Words are also irrelevant to the events, because the event builds on top of the concepts. Recalling that the corresponding rows of W_a or W_t are identically equal to zero for the irrelevant or noisy features, we should be able to find similar patterns in the distribution of these rows 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 $W = [w^1, ..., w^{d^h}]^T$ where w^i is the vertical 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 [7][26][25]. The sparse models are used to make some of the feature coefficients shrink to zeros to achieve feature selection. The "shrinking to zero" idea can be applied to uncover the common distribution of the "identically equal to zero" rows of W_a and W_t discussed before. In this way, we can remove the shared irrelevance and noise, thus obtaining a more discriminative W_t.

Now we introduce the technical details of our joint sparsity model. Specifically, we propose to exploit $\|W\|_{2,p} = \left(\sum_{i=1}^{d_h}(\sum_{j=1}^{c_a+1}|W_{ij}|)^{\frac{p}{2}}\right)^{\frac{1}{p}}$ to achieve that goal. $\|\cdot\|_{2,p}$ denotes the $\ell_{2,p}$-norm $(0 < p < 2)$. By minimizing $\|W\|_{2,p}^p$, we can reduce the negative impact of the irrelevant or noisy w_i's. Our model has the flexibility of characterizing different degree of relevance between concepts and events. p is used to control the degree of shared structures. The lower p is, the more semantically correlated are the auxiliary concepts and the target event. By contrast, when the auxiliary concepts and the target event have less relevance, we can use a larger p. When we increase p to 2, we do not impose sharing on the two domains. To step further, it is expected that the predicted labels of W_t on $\tilde{X}_t$ be consistent with those of P_t on $\tilde{Z}_t$, thus resulting in more accurate P_t and W_t. In this way, P_t from the heterogeneous features of the target and W_t from the knowledge adaptation would jointly augment the observations for MED. We achieve this by minimizing $\left\|\tilde{X}_t^T W_t - \tilde{Z}_t^T P_t\right\|_F^2$ where $\|\cdot\|_F^2$ indicates the Frobenius norm of a matrix.

To this end, we propose the following objective function for MED with few exemplars:

$$\begin{aligned}\min_{P_t,W_t,W_a,b_t,b_a} & \left\|\tilde{Z}_t^T P_t + 1_t b_t - y_t\right\|_{2,1} + \left\|\tilde{X}_t^T W_t - \tilde{Z}_t^T P_t\right\|_F^2 \\ & + \left\|\tilde{X}_a^T W_a + 1_a b_a - Y_a\right\|_{2,1} + \alpha\|W\|_{2,p}^p + \beta(\|W_t\|_F^2 + \|W_a\|_F^2)\end{aligned} \tag{5}$$

where $(\|W_a\|_F^2 + \|W_t\|_F^2)$ is added to avoid over-fitting. α and β are regularization parameters.

Once P_t and W_t are obtained, we apply them to the testing videos for event detection. The decision values of them are normalized and then their weighted sum based on the feature numbers are the final decision values of the testing videos. Our method builds upon 1) the knowledge adaptation from concepts-based videos to event-based videos by leveraging the shared structures between them; and 2) the augmented observation from the particular features that are only owned by MED videos. We therefore name our method Heterogenous Features based Structural Adaptive Regression (HF-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.

Denote $\tilde{Z}_t^T P_t - y_t = [u^1, ..., u^{n_t}]^T$, $\tilde{X}_a^T W_a - Y_a = [v^1, ..., v^{n_a}]^T$. Next, we define three diagonal matrices D_t, D_a and D with their diagonal elements $D_t^{ii} = \frac{1}{2\|u^i\|_2}$, $D_a^{ii} = \frac{1}{2\|v^i\|_2}$, $D^{ii} = \frac{1}{\frac{2}{p}\|w^i\|_2^{2-p}}$ respectively. The objective function in Eq. (5) is equivalent to:

$$\begin{aligned}
&\min_{\substack{P_t, W_t, W_a, \\ b_t, b_a}} Tr\left((\tilde{Z}_t^T P_t + 1_t b_t - y_t)^T D_t (\tilde{Z}_t^T P_t + 1_t b_t - y_t)\right) \\
&+ \left\|\tilde{X}_t^T W_t - \tilde{Z}_t^T P_t\right\|_F^2 + Tr\left((\tilde{X}_a^T W_a + 1_a b_a - Y_a)^T D_a \right. \\
&\left. (\tilde{X}_a^T W_a + 1_a b_a - Y_a)\right) + \alpha Tr\left(W^T D W\right) \\
&+ \beta(\|W_a\|_F^2 + \|W_t\|_F^2)
\end{aligned} \tag{6}$$

where $Tr(\cdot)$ denotes the trace operator. By setting the derivative of Eq. (6) *w.r.t.* b_a to zero, we get:

$$b_a = \frac{1}{n_a} 1_a^T Y_a - \frac{1}{n_a} 1_a^T \tilde{X}_a^T W_a. \tag{7}$$

Similarly, we obtain b_t as:

$$b_t = \frac{1}{n_t} 1_t^T y_t - \frac{1}{n_t} 1_t^T \tilde{Z}_t^T P_t. \tag{8}$$

Substituting Eq. (7) and Eq. (8) into Eq. (6), it becomes:

$$\begin{aligned}
&\min_{P_t, W_t, W_a} Tr\left((H_t \tilde{Z}_t^T P_t - H_t y_t)^T D_t (H_t \tilde{Z}_t^T P_t - H_t y_t)\right) \\
&+ \left\|\tilde{X}_t^T W_t - \tilde{Z}_t^T P_t\right\|_F^2 \\
&+ Tr\left((H_a \tilde{X}_a^T W_a - H_a Y_a)^T D_a (H_a \tilde{X}_a^T W_a - H_a Y_a)\right) \\
&+ \alpha Tr\left(W^T D W\right) + \beta(\|W_a\|_F^2 + \|W_t\|_F^2)
\end{aligned} \tag{9}$$

where $H_t = I_t - \frac{1}{n_t} 1_t 1_t^T$, $H_a = I_a - \frac{1}{n_a} 1_a 1_a^T$ and $I_t \in \mathbb{R}^{n_t \times n_t}$, $I_a \in \mathbb{R}^{n_a \times n_a}$ are two identity matrices. Setting the derivative of Eq. (9) *w.r.t.* W_a to zero, we get:

$$W_a = (\tilde{X}_a H_a D_a H_a \tilde{X}_a^T + \alpha D + \beta I_d)^{-1} \tilde{X}_a H_a D_a H_a Y_a \tag{10}$$

where $I_d \in \mathbb{R}^{d_h \times d_h}$ is an identity matrix. Note that D is treated as a constant in this step as we adopt an alternating optimization approach here. In the same manner, we obtain the event detector W_t as:

$$W_t = A^{-1} \tilde{X}_t \tilde{Z}_t P_t \tag{11}$$

where $A = \alpha D + \beta I_d + \tilde{X}_t \tilde{X}_t^T$. To optimize P_t, the problem equals to:

$$\begin{aligned}
&\min_{P_t} Tr(P_t^T \tilde{Z}_t H_t D_t H_t \tilde{Z}_t^T P_t - 2 P_t^T \tilde{Z}_t H_t D_t H_t y_t) \\
&+ \left\|\tilde{X}_t^T W_t - \tilde{Z}_t^T P_t\right\|_F^2 + \alpha Tr(W_t^T D W_t) + \beta Tr(W_t^T W_t)
\end{aligned} \tag{12}$$

Substituting Eq. (11) into Eq. (12) and defining

$$J = \tilde{Z}_t H_t D_t H_t \tilde{Z}_t^T - \tilde{Z}_t \tilde{X}_t^T A^{-1} \tilde{X}_t \tilde{Z}_t^T + \tilde{Z}_t \tilde{Z}_t^T \tag{13}$$

$$K = 2\tilde{Z}_t H_t D_t H_t y_t, \tag{14}$$

the problem becomes:

$$\min_{P_t} Tr(P_t^T J P_t - P_t^T K) \tag{15}$$

By setting the derivative of the above function *w.r.t.* P_t to zero, we get:

$$P_t = \frac{1}{2} J^{-1} K \tag{16}$$

Next, we propose Algorithm 1 to solve the objective function in Eq. (5). The computational complexity of Algorithm 1 is as follows. For training, it is $\mathcal{O}(d_z^3)$ as $d_z > d_h$. Note that $d_z \gg n_t$ because there are few training examples in our problem. 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_z dimensional so the complexity is $\mathcal{O}(d_z n_t n_{te})$.

It can be proved that the objective function value of Eq. (5) monotonically decreases in each iteration until converging to local optimum using Algorithm 1.

6 Experiments

In this section, we present the experiments which evaluate the performance of our Heterogenous Features based Structural Adaptive Regression (HF-SAR) for MED with few exemplars.

6.1 Datasets

NIST has provided so far the largest video corpora for MED. Our experiments on MED with few exemplars 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

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

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

Algorithm 1: Optimizing the event detector.

```
Input:
    The target training data $\tilde{Z}_t \in \mathbb{R}^{d_z \times n_t}$, $\tilde{X}_t \in \mathbb{R}^{d_h \times n_t}$, $y_t \in \mathbb{R}^{n_t \times 1}$;
    The auxiliary data $\tilde{X}_a \in \mathbb{R}^{d_h \times n_a}, Y_a \in \mathbb{R}^{n_a \times c_a}$;
    Parameters $\alpha$, $\beta$ and $p$.
Output:
    Optimized $P_t \in \mathbb{R}^{d_z \times 1}$, $W_t \in \mathbb{R}^{d_h \times 1}$ and $b_t \in \mathbb{R}^1$.
Set $t = 0$, initialize $P_t \in \mathbb{R}^{d_z \times 1}$, $W_t \in \mathbb{R}^{d_h \times 1}$ and $W_a \in \mathbb{R}^{d_h \times c_a}$ randomly;
repeat
    Compute $\tilde{Z}_t^T P_t - y_t = [u^1, ..., u^{n_t}]^T$, $\tilde{X}_a^T W_a - Y_a = [v^1, ..., v^{n_a}]^T$ and $W = [w^1, ..., w^d]^T$;
    Compute the diagonal matrix $D_t^t$, $D_a^t$ and $D^t$ according to $D_t^{ii} = \frac{1}{2\|u^i\|_2}$, $D_a^{ii} = \frac{1}{2\|v^i\|_2}$, and $D^{ii} = \frac{1}{\frac{2}{p}\|w^i\|_2^{2-p}}$ respectively;
    Update $W_a^{t+1}$ as: $W_a^{t+1} = (\tilde{X}_a H_a D_a^t \tilde{X}_a^T + \alpha D^t + \beta I_d)^{-1} \tilde{X}_a H_a D_a H_a Y_a^T$;
    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 $P_t^{t+1}$ according to Eq. (13), Eq. (14) and Eq. (16);
    Update $W_t^{t+1}$ as: $W_t^{t+1} = A^{-1} \tilde{X}_t \tilde{Z}_t^T P_t^{t+1}$;
    Update $b_t^{t+1}$ as: $b_t^{t+1} = \frac{1}{n_t} 1_t^T y_t - \frac{1}{n_t} 1_t^T \tilde{Z}_t^T W_t^{t+1}$;
    $t = t + 1$.
until Convergence;
Return $P_t$, $W_t$ and $b_t$.
```

Working on a sewing project. The two datasets are combined together (MED10-11 for short) in our experiments so we have a dataset of 9746 video clips.

We first use the development set from TRECVID 2012 semantic indexing task (SIN12) as the auxiliary videos. SIN12 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" [15]. 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 3244 video frames. On the other hand, multimedia events are usually accompanied by human actions, which suggests that we may find similar motion features between event videos and basic human action videos. Hence, we additionally use UCF50 dataset [16] to test whether it is able to facilitate multimedia event detection.

6.2 Setup

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 for all the videos. When utilizing SIN12 dataset, we extract SIFT [23] and CSIFT [34] features for the videos in MED10-11 and SIN12. Then we use 1x1, 2x2 and 3x1 spatial grids to generate the spatial BoW representation [35]. For each grid, we use a standard BoW representation with 4,096 dimensions, thus resulting in a 32,768 dimension spatial BoW feature for SIFT/CSIFT to represent each video. When utilizing UCF50 dataset, we extract STIP [36] feature for the videos in MED10-11 and UCF50 since STIP has proved to be robust for analyzing action videos. A similar procedure is followed to generate the spatial BoW representation. Hence, the feature representation in this work is different from our previous work [13] in which we just used BoW representation. Apart from visual features, some other features, which provide different yet complementary information, can also be used to represent videos. For example, the auditory feature based on Mel-frequency Cepstral Coefficients (MFCC) has also been frequently used [19]. We additionally use this feature for MED videos and the dimension is 4096. Thus, when using the SIN12 dataset, our two domains have SIFT and CSIFT as shared feature type while MFCC works as the heterogeneous feature for MED videos; when using UCF50 dataset, our two domains have STIP as shared feature type while MFCC is the heterogeneous feature for MED videos.

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 MED with few exemplars [2]. However, for MED10-11 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 definition given by NIST and randomly select 10 positive examples for each event. Other 1000 negative examples are selected and combined with the positive examples as the training data. The remaining 8736 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 as 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 Comparison Algorithms

In this section, we show the MED results using Heterogenous Features based Structural Adaptive Regression (HF-SAR) and other state-of-the-art algorithms. A brief introduction of the comparison algorithms is as follows:

- HF-SAR: the proposed new method which is designed for knowledge adaptation based on heterogeneous features. The χ^2 kernel is used for its advantageous performance on video analysis.
- Structural Adaptive Regression (SAR) [13]: our previous algorithm on knowledge adaptation for MED with few exemplars. Similarly, the χ^2 kernel is used.
- Adaptive Multiple Kernel Learning (A-MKL) [12]: a recent knowledge adaptation algorithm built upon SVM.
- Multiple Kernel Transfer Learning (MKTL) [14]: a recent multi-class transfer learning algorithm built within a multiple kernel learning framework. The original algorithm in [14] has used RBF kernel. For fair comparison, we implement it with χ^2 kernel.
- SAR&SVM: We use SAR based on SIFT+CSIFT features between the auxiliary domain and the target domain. In addition, we use SVM based on MFCC feature in the target domain. Then we fuse the decision values obtained by both of them. In this way, we can evaluate the performance of combining homogeneous transfer learning and the classifier on the heterogeneous feature.
- SVM: the most widely used and robust event detector for MED [19][10][17][37]. Similarly, we use the χ^2 kernel for it.
- TaylorBoost [28]: a state-of-the-art classifier extended from AdaBoost.

For SVM, we use LIBSVM, and for A-MKL, MKTL and TaylorBoost we use the code shared by the authors. During the training and predicting, we combine SIFT, CSIFT and MFCC features of the MED10-11 dataset for SVM and TaylorBoost. SAR, A-MKL and MKTL are knowledge adaptation based algorithms, which utilize the SIN12 dataset as auxiliary data. However, they require that the target domain and the auxiliary domain have the homogeneous feature representation so only SIFT and CSIFT are used for them. HF-SAR leverages SIN12 for MED with few exemplars on MED10-11 and it is capable of using SIFT, CSIFT and MFCC together.

All the regularization parameters are tuned from $\{0.001, 0.1, 10, 1000\}$, and the parameter p of HF-SAR and SAR is tuned from $\{0.5, 1, 1.5\}$. We report the best results for each algorithm.

6.4 MED Results

The detection performance of different algorithms is displayed in Figure 3 and Table 1 where all the knowledge adaptation methods have exploited SIN12 dataset. Note that LOWER MinNDC and Pmd@TER=12.5 indicate BETTER performance; HIGHER AP indicates BETTER performance. The proposed method HF-SAR is consistently competitive. Zooming into details, we have the following observations: 1) when using MinNDC as metric, HF-SAR gains the best performance for 17 events; 2) when using Pmd@TER=12.5 as metric, HF-SAR gains the best performance for 15 events; 3) when using AP as metric, HF-SAR is the best method for 14 events; 4) HF-SAR obtains the top performance for the average accuracy over all the 18 events; 5) SAR&SVM is generally the second competitive algorithm. This indicates that incorporating the additional information contained in the heterogenous feature into a robust knowledge adaptation algorithm based on homogeneous features is beneficial. However, it is unclear which algorithms should be combined together for the best performance as they may work with different mechanisms; 6) SAR, A-MKL and SVM have varying degrees of success on some events. However, they are generally worse than HF-SAR and SAR&SVM. It means knowledge adaptation based on homogeneous features loses useful information from the heterogenous feature, and SVM utilizes all the features but it cannot leverage knowledge from other sources. In contrast, the newly proposed method HF-SAR transfers knowledge between homogeneous features while simultaneously exploits the heterogeneous feature to get boosted performance.

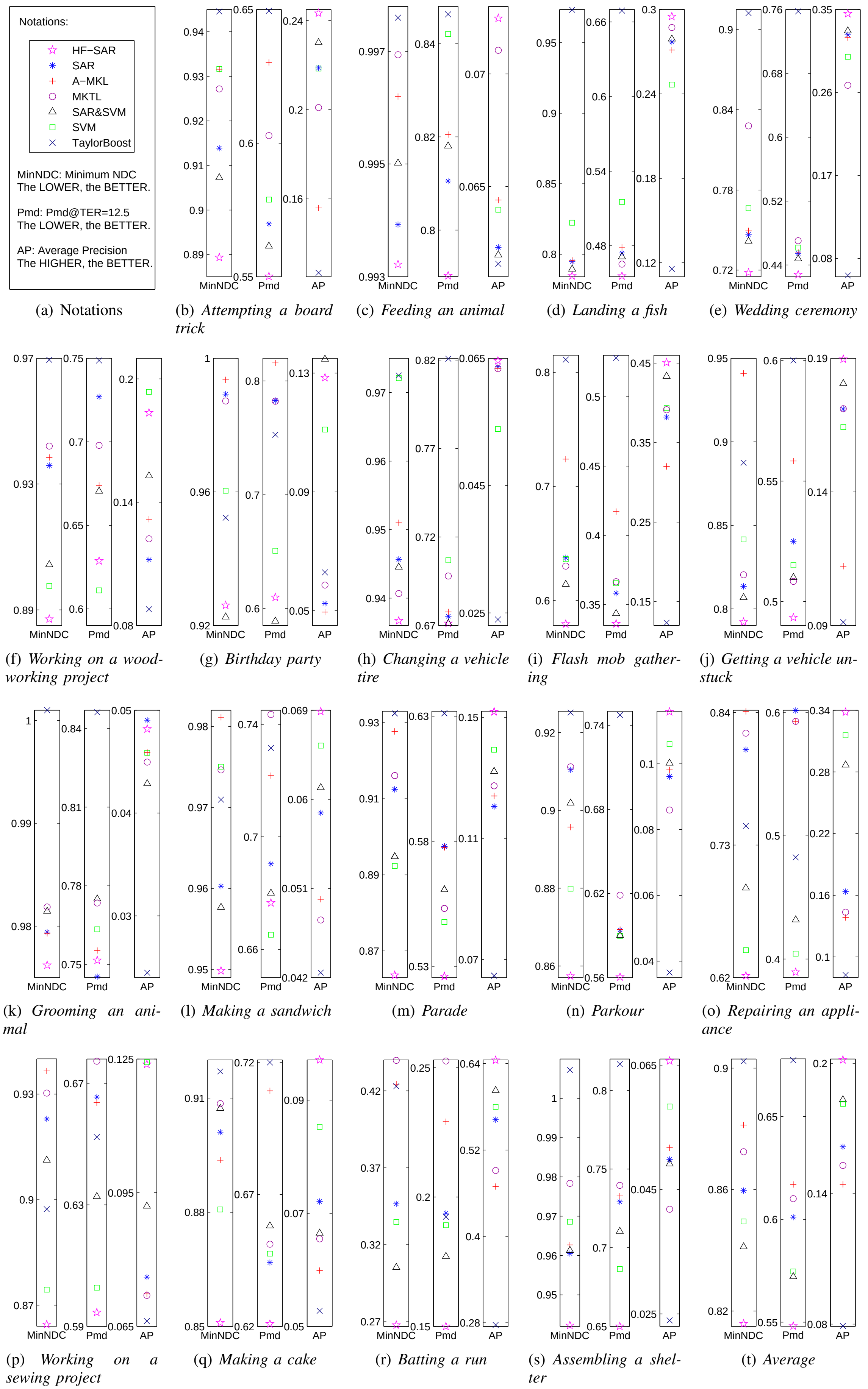
Next we show the detection results by exploiting UCF50 dataset. As HF-SAR has already shown its advantage over other knowledge adaptation algorithms and this experiment aims to show that we can even adapt useful action knowledge for MED with few exemplars, we only compare HF-SAR to the best baseline classifier SVM. This time, we combine STIP and MFCC features of the MED10-11 dataset for SVM. The detailed results are illustrated in Table 2. As can be seen, HF-SAR beats SVM on 17, 17, 15 events and the average performance over all the 18 events in terms of MinNDC, Pmd@TER=12.5, AP respectively. Moreover, for those events on which HF-SAR is better, we can observe noticeable performance improvement. It is also worth mentioning that the performance of SVM is lower than its performance in the previous experiment because different and fewer features are used.

6.5 Influence of Knowledge Adaptation

It is interesting to understand how the knowledge adaptation from the auxiliary concept-based videos impacts the MED with few exemplars. We base our study on two scenarios: First, we set α in Eq. (5) to 0 so there is no knowledge adaptation; Second, since in our objective function the item $\alpha \|W\|_{2,p}^{p}$ controls the effect of the knowledge adaptation, we investigate the influence by varying the parameter α and p after fixing β at its optimal values.

For the first scenario, we show the performance comparison between using auxiliary data and not using it in Figure 4. MinNDC is used as metric and the results on the first 10 events are displayed due to the space limit. It can be seen that using auxiliary data has clear advantage over not using it, which demonstrates that through proper design, the auxiliary knowledge contributes notably to the MED with few exemplars.

For the second scenario, we similarly use MinNDC as metric to show the performance variation. Due to the space limit, we only show the results on the first 6 events in Figure 5. We observe from Figure 5 that the best results are generally obtained when $p = 0.5$ or $p = 1$. For the other parameter α there is no obvious rule, which is presumably data-dependent. Lower p indicates that the model is more sparse, thereby eliminating more redundancy and noise.



(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 MED with few exemplars.

Table 1
Average detection accuracy of different methods. Better results are highlighted in bold.

Evaluation Metric	SAR	A-MKL	MKTL	SAR&SVM	SVM	TaylorBoost	HF-SAR
MinNDC	0.860	0.881	0.873	0.841	0.850	0.902	**0.817**
Pmd@12.5	0.601	0.617	0.610	0.572	0.575	0.677	**0.549**
AP	0.162	0.144	0.153	0.183	0.181	0.080	**0.201**

Table 2
Detection results by exploiting UCF50 dataset in comparison with SVM.

Event	Metric	SVM	HF-SAR	Relative Improvement
E01	MinNDC	0.884	0.922	N/A
	Pmd@TER=12.5	0.546	0.569	N/A
	AP	0.247	0.206	N/A
E02	MinNDC	1.000	0.999	0.1%
	Pmd@TER=12.5	0.938	0.877	7.0%
	AP	0.037	0.046	24.3%
E03	MinNDC	1.000	0.990	1.0%
	Pmd@TER=12.5	0.928	0.815	13.9%
	AP	0.035	0.061	74.3%
E04	MinNDC	0.936	0.912	2.6%
	Pmd@TER=12.5	0.870	0.770	13.0%
	AP	0.044	0.132	200%
E05	MinNDC	0.975	0.946	3.1%
	Pmd@TER=12.5	0.914	0.817	11.9%
	AP	0.061	0.097	59.0%
E06	MinNDC	0.992	0.973	2.0%
	Pmd@TER=12.5	0.917	0.797	15.1%
	AP	0.049	0.077	57.1%
E07	MinNDC	1.000	0.992	0.8%
	Pmd@TER=12.5	0.944	0.881	7.2%
	AP	0.033	0.032	N/A
E08	MinNDC	0.945	0.833	13.4%
	Pmd@TER=12.5	0.862	0.676	27.5%
	AP	0.094	0.173	84.0%
E09	MinNDC	0.970	0.928	4.5%
	Pmd@TER=12.5	0.804	0.703	14.4%
	AP	0.072	0.093	29.2%
E10	MinNDC	0.997	0.991	0.6%
	Pmd@TER=12.5	0.933	0.862	8.2%
	AP	0.035	0.043	22.9%
E11	MinNDC	0.995	0.982	1.3%
	Pmd@TER=12.5	0.904	0.835	8.3%
	AP	0.037	0.041	10.8%
E12	MinNDC	0.975	0.940	9.4%
	Pmd@TER=12.5	0.889	0.770	4.5%
	AP	0.052	0.077	13.7%
E13	MinNDC	0.970	0.957	3.7%
	Pmd@TER=12.5	0.711	0.689	3.2%
	AP	0.094	0.078	N/A
E14	MinNDC	0.919	0.819	12.2%
	Pmd@TER=12.5	0.840	0.687	22.3%
	AP	0.083	0.191	130.1%
E15	MinNDC	0.964	0.945	2.0%
	Pmd@TER=12.5	0.880	0.794	10.8%
	AP	0.059	0.066	11.9%
E16	MinNDC	0.975	0.937	4.1%
	Pmd@TER=12.5	0.864	0.796	8.5%
	AP	0.045	0.053	17.8%
E17	MinNDC	0.893	0.736	21.3%
	Pmd@TER=12.5	0.766	0.585	30.9%
	AP	0.125	0.253	102.4%
E18	MinNDC	0.982	0.967	1.6%
	Pmd@TER=12.5	0.922	0.836	10.3%
	AP	0.036	0.041	13.9%
Average	MinNDC	0.965	0.932	3.5%
	Pmd@TER=12.5	0.857	0.764	12.2%
	AP	0.069	0.098	42.0%

6.6 Using Fewer Concepts

In this experiment, we test the performance variance of the proposed algorithm by varying the number of auxiliary concepts as 5, 10, 20, 30, 50 and 65. Figure 6 displays the corresponding results on the first 10 events in terms of Minimum NDC. We have the following observations: 1) Generally, the performance of using only 5 auxiliary concepts is noticeably worse than using all the 65 auxiliary concepts; 2) From using 5 auxiliary concepts to using 30 auxiliary concepts, the performance is gradually improved for most events; 3) From using 30 auxiliary concepts to using 65 auxiliary concepts, the performance does not vary much, which suggests that the performance saturates at the point when 30 auxiliary concepts are used. Our observation indicates that when the number of auxiliary concepts is very small, which also means few auxiliary videos, the performance gain is limited. To get more performance boost, we may want to incorporate more auxiliary videos with more concepts. However, how to decide the optimal number of auxiliary concepts is still an open problem and is out of the scope of this paper.

6.7 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 500 and 100. Figure 7 shows the performance comparison between using 100, 500 and 1000 negative examples on the first 10 events. Similarly, Minimum NDC is chosen as the evaluation metric.

From Figure 7 we have the following observations: 1) Using 1000 or 500 negative examples is better than using only 100 negative examples. 2) The performance difference between using 1000 and using 500 negative examples is quite small. This experiment indicates that negative examples are helpful in improving the detection accuracy in some degree. For example, when 500 or 1000 negative examples are used, the performance is generally better than using 100 negative examples only. However, as the number of negative examples used increases, the performance gain does not increase accordingly, *e.g.*, from using 500 negative examples to using 1000 negative examples. How many negative examples would bring in the most performance gain is still an open problem, which is out of the scope of this paper. However, since negative examples are quite easy to obtain in the real world, it is natural to leverage such cheap resources for boosted detection accuracy.

6.8 Parameter Sensitivity Study

There are two regularization parameters, denoted as α and β in Eq. (5). To learn how they affect the performance, we conduct an experiment on the parameter sensitivity. Due to the space limit, we still only show the results on the first 6

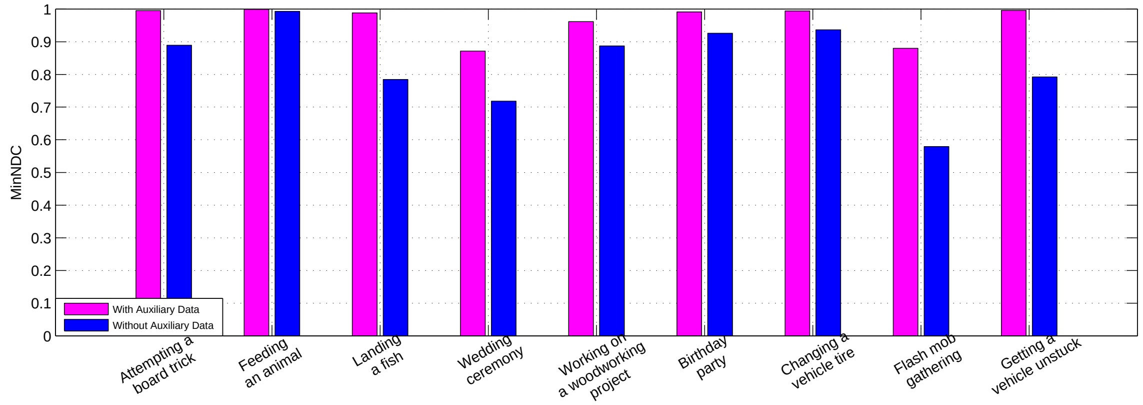
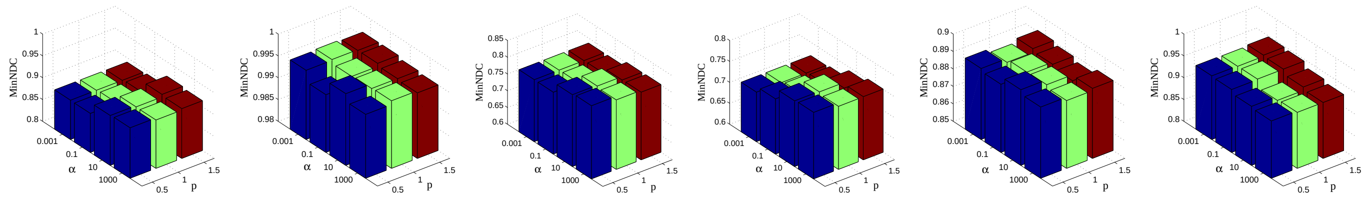


Figure 4. Performance comparison between using auxiliary knowledge and not using auxiliary knowledge.



(a) *Attempting a board trick* (b) *Feeding an animal* (c) *Landing a fish* (d) *Wedding ceremony* (e) *Working on a woodworking project* (f) *Birthday party*

Figure 5. The detection performance variance *w.r.t.* α and p.

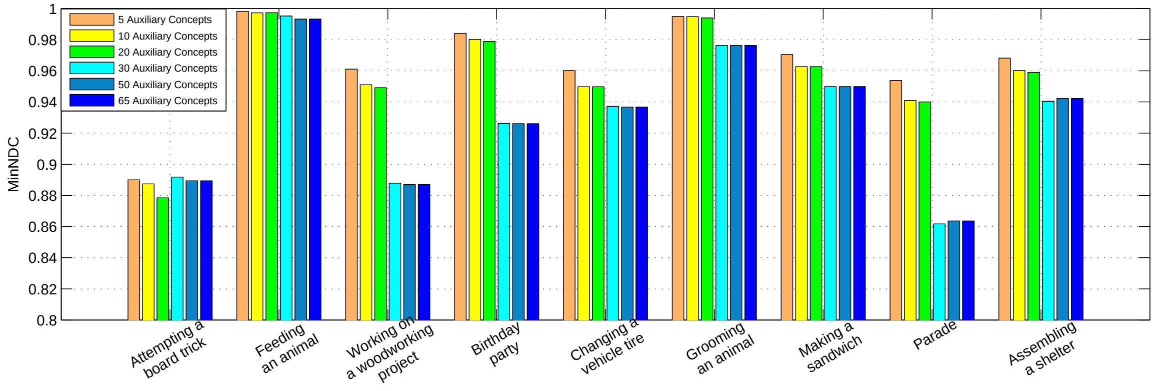


Figure 6. Performance comparison between using 5, 10, 20, 30, 50 and 65 auxiliary concepts.

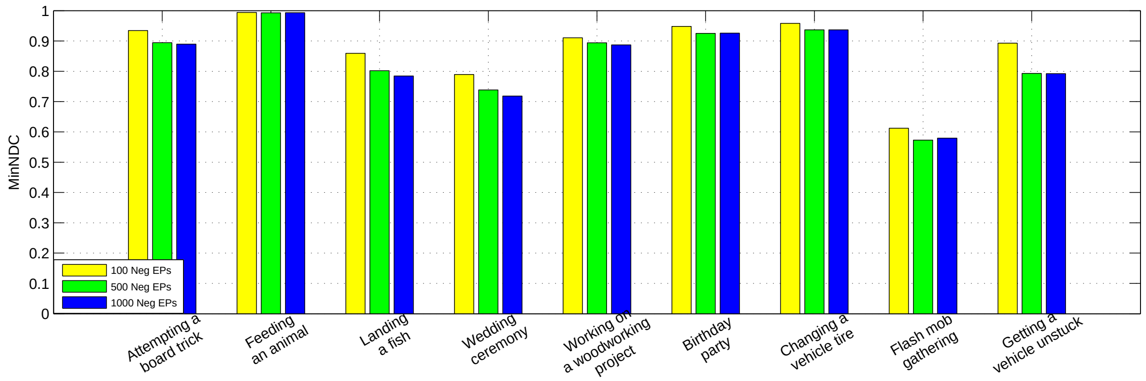
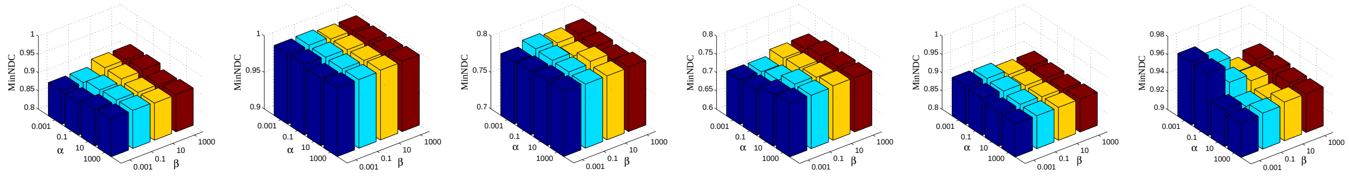


Figure 7. Performance comparison between using 100, 500 and 1000 negative examples.



(a) *Attempting a board trick* (b) *Feeding an animal* (c) *Landing a fish* (d) *Wedding ceremony* (e) *Working on a woodworking project* (f) *Birthday party*

Figure 8. The detection performance variance *w.r.t.* α and β.

events in Figure 8. From Figure 8 we notice that for some events, *e.g.*, *Birthday party*, the performance is sensitive to the two parameters. For some other events like *Feeding an animal* the performance does not change much. However, we can generally obtain good performance for these events when α and β are comparable. For example, good performance is obtained when $\alpha = \beta = 0.001$ for *Attempting a board trick*, *Feeding an animal*, *Landing a fish* and *Wedding ceremony*, and $\alpha = \beta = 10$ for *Birthday party*. A similar pattern is observed for other events as well.

6.9 Convergence Study

We solve our objective function using an alternating approach. In practice, how fast our algorithm converges is crucial for the whole computational efficiency. Hence, we conduct an experiment to show the convergence curve of our algorithm. As we have similar results on all the 18 events, we only present the convergence curve on the first event. All the parameters involved are fixed at 1.

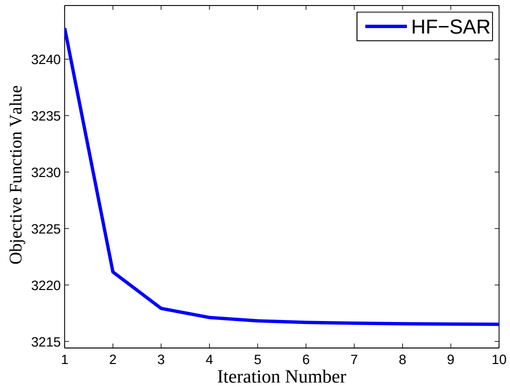


Figure 9. Convergence curve of the objective function value in Eq. (5) using Algorithm 1 for the event *Attempting a board trick*. The figure shows that the objective function value monotonically decreases until convergence by applying the proposed algorithm.

Figure 9 shows the convergence curve. It can be seen that the objective function value converges within 10 iterations. The convergence experiment demonstrates the efficiency of our alternating algorithm.

7 Complementary Experiment on Multi-Class Classification

Our proposed algorithm can be easily extended to other applications such as multi-class classification. In this section, we conduct a complementary experiment on image annotation to show its effectiveness for multi-class classification.

We use the Animals with Attributes (AwA) dataset [42] for evaluation. The reason is that the dataset has both animal categories and the associated attributes. Similarly to our assumption, different animal categories may share common attributes. Thus, we use the 10 animal categories specified in [42] as our target annotation categories and the rest as our auxiliary data. Note that for the auxiliary data we use their attribute labels since these attributes are the shared components with the target animal categories. The 10 target categories are *persian cat*, *hippopotamus*, *leopard*, *humpback whale*, *seal*, *chimpanzee*, *rat*, *giant panda*, *pig* and *raccoon*. For the 10 classes to be annotated, we randomly select 10 samples per category to form the training set and the remaining data of these categories are our testing data. We use the SIFT feature as the homogeneous feature and the Locality Similarity Histogram (LSH) feature as the heterogeneous feature for image representation. In other words, the images of the 10 target categories are represented by SIFT and LSH while those of the auxiliary categories are represented by SIFT only.

The annotation comparison between different algorithms is displayed in Table 3. We can see that HF-SAR is much better than other comparison algorithms. SVM is second best algorithm. Especially, other transfer learning algorithms have weaker performance as only one feature is exploited.

The reported accuracy in [42] is 40.5%. But we point out that in [42] six features have been used whereas we only use two features. We did not use all the features used in [42] because we were concerned with the computational efficiency, *e.g.*, the comparison algorithms A-MKL and TaylorBoost are computationally expensive. On the other hand, to be consistent with our previous experiment on MED with few exemplars, we select 10 samples from each target category to form the training set, making our training set also different from that in [42]. Thus, we cannot directly compare the annotation accuracy of our method and that of [42].

This complementary experiment has demonstrated that our method also has the potential for other applications.

8 Conclusion

In this paper, we have introduced the research exploration of MED with few exemplars. This is an important research issue

Table 3
Performance comparison of different methods on image annotation. The best result is highlighted in bold.

Evaluation Metric	SAR	A-MKL	MKTL	SAR&SVM	SVM	TaylorBoost	HF-SAR
Accuracy	0.257	0.248	0.232	0.265	0.310	0.264	**0.373**

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 MED with few exemplars. A notable difference between our proposed algorithm and most existing knowledge adaptation algorithms is that it is built upon heterogeneous features, *i.e.*, the features of the source and the target are partially different, but overlapping. Specifically, we first mine the shared irrelevance and noise between the auxiliary videos and the target videos based on the homogeneous features. Then a sophisticated method is exerted to alleviate the negative impact of the irrelevance and noise to optimize the event detector. Meanwhile, another event detector of MED videos is trained based on the heterogeneous feature. Then we integrate the two event detectors for optimization, after which the decision values from both are fused for the final prediction. Extensive experiments using real-world multimedia archives were conducted with promising results. The results validate that: 1) it is beneficial to leverage auxiliary knowledge for MED when we do not have sufficient positive examples; and 2) the capability of knowledge adaptation based on heterogeneous features is realistic and advantageous.

9 Acknowledgments

This paper was partially supported by the European Commission under contract FP7-611346 xLiMe, the US Department of Defense the U. S. Army Research Office (W911NF-13-1-0277) and the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center contract number D11PC20068. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, ARO, or the U.S. Government.

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] G. Zen, and E. Ricci. Earth mover's prototypes: A convex learning approach for discovering activity patterns in dynamic scenes. In *CVPR*, pages 3225–3232, 2011.

[6] E. Ricci, G. Zen, N. Sebe, and S. Messelodi. A prototype learning framework using EMD: Application to complex scenes analysis. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(3): 513–526, 2013.

[7] A. Argyriou, T. Evgeniou, and M. Pontil. Convex multi-task feature learning. *Machine Learning* (73): 243–272, 2008.

[8] G. Obozinski, B. Taskar, and M. I. Jordan. Joint covariate selection and joint subspace selection for multiple classification problems. *Statistics and Computing* (20): 231–252, 2010.

[9] Y. Yang, Z. Ma, A. G. Hauptmann, and N. Sebe. Feature Selection for Multimedia Analysis by Sharing Information Among Multiple Tasks. *IEEE Transactions on Multimedia*, 15(3): 661–669, 2013.

[10] 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. G. Hauptmann. Informedia @ TRECVID2011. In *NIST TRECVID Workshop*, 2011.

[11] 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.

[12] L. Duan, D. Xu, I. W.-H. Tsang, and J. Luo. Visual event recognition in videos by learning from web data. *IEEE Trans. Pattern Anal. Mach. Intell.*, 34(9): 1667–1680, 2012.

[13] 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*, pages 469–478, 2012.

[14] J. Luo, T. Tommasi, and B. Caputo. Multiclass transfer learning from unconstrained priors. In *ICCV*, pages 1863–1870, 2011.

[15] 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*, 285: 157–171, 2010.

[16] K. K. Reddy and M. Shah. Recognizing 50 human action categories of web videos. *Machine Vision and Applications Journal*, 24(5): 971–981, 2013.

[17] 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.

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

[19] Z.-Z. Lan, L. Bao, S.-I. Yu, W. Liu, and A. G. Hauptmann. Multimedia Classification and Event Detection using Double Fusion. *Journal of Multimedia Tools and Applications*, 2013.

[20] Z. Ma, Y. Yang, Z. Xu, S. Yan, N. Sebe, and A. G. Hauptmann. Complex Event Detection via Multi-source Video Attributes. In *CVPR*, pages 2627–2633, 2013.

[21] Y. Yang, Z. Ma, Z. Xu, S. Yan, and A. G. Hauptmann. How Related Exemplars Help Complex Event Detection inWeb Videos. In *ICCV*, 2013.

[22] 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.

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

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

[25] Z. Ma, F. Nie, Y. Yang, J. R. R. Uijlings, and N. Sebe. Web image annotation via subspace-sparsity collaborated feature selection. *IEEE Transactions on Multimedia*, 14(4): 1021–1030, 2012.

[26] 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 MM*, pages 283–292, 2011.

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

[28] 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.

[29] 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.

[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] 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.

[32] C. G. M. 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 MM*, pages 421–430, 2006.

[33] 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.

[34] 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.

[35] M. Marszalek, C. Schmid, H. Harzallah, and J. van de Weijer. Learning object representations for visual object class recognition. In *Visual recognition challenge workshop*, 2007.

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

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

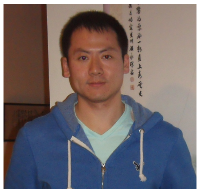
[38] G. Wang, T.-S. Chua, and M. Zhao. Exploring knowledge of sub-domain in a multi-resolution bootstrapping framework for concept detection in news video. In *ACM MM*, pages 249–258, 2008.

[39] 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 MM*, pages 221–230, 2006.

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

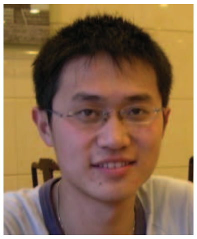
[41] Y. Yang, F. Nie, D. Xu, J. Luo, Y. Zhuang, and Y. Pan. A multimedia retrieval framework based on semi-supervised ranking and relevance feedback. *IEEE Trans. Pattern Anal. Mach. Intell.*, 34(4): 723–742, 2012.

[42] C. H. Lampert, H. Nickisch and S. Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *CVPR*, pages 951–958, 2009.



Zhigang Ma received the Ph.D. in computer science from University of Trento, Trento, Italy, in 2013.

He is now a Postdoctoral Research Fellow with the School of Computer Science, Carnegie Mellon University, Pittsburgh, PA. His research interest is mainly on machine learning and its applications to multimedia analysis and computer vision.



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

He is now a DECRA fellow with the University of Queensland, Brisbane, Australia. Prior to that, he was 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, surveillance video analysis, 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 a General Co-Chair of the IEEE Automatic Face and Gesture Recognition Conference, FG 2008 and ACM Multimedia 2013, and a program chair of ACM International Conference on Image and Video Retrieval (CIVR) 2007 and 2010, ACM Multimedia 2007 and 2011. He is a program chair of ECCV 2016 and ICCV 2017. He is a senior member of IEEE and of ACM and a fellow of IAPR.

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.