How Related Exemplars Help Complex Event Detection in Web Videos?

Yi Yang[§†] Zhigang Ma[§] Zhongwen Xu[†] Shuicheng Yan[‡] Alexander G. Hauptmann[§]
[§]School of Computer Science, Carnegie Mellon University, USA
[†]ITEE, The University of Queensland, Australia
[‡]ECE, National University of Singapore, Singapore
{yiyang,kevinma,alex}@cs.cmu.edu z.xu3@uq.edu.au eleyans@nus.edu.sg

Abstract

*Compared to visual concepts such as actions, scenes and objects, complex event is a higher level abstraction of longer video sequences. For example, a "marriage proposal" event is described by multiple objects (e.g., ring, faces), scenes (e.g., in a restaurant, outdoor) and actions (e.g., kneeling down). The positive exemplars which exactly convey the precise semantic of an event are hard to obtain. It would be beneficial to utilize the related exemplars for complex event detection. However, the semantic correlations between related exemplars and the target event vary substantially as relatedness assessment is subjective. Two related exemplars can be about completely different events, e.g., in the TRECVID MED dataset, both bicycle riding and equestrianism are labeled as related to "*attempting a bike trick*" event. To tackle the subjectiveness of human assessment, our algorithm automatically evaluates how positive the related exemplars are for the detection of an event and uses them on an exemplar-specific basis. Experiments demonstrate that our algorithm is able to utilize related exemplars adaptively, and the algorithm gains good performance for complex event detection.*

1. Introduction

Current research of visual content analysis mainly focuses on the recognition of visual concepts, such as actions, scenes and objects [1][2] [17]. Differently, in this paper we propose a generic framework to detect complex events in large scale unstructured web video archives. Figure 1 shows two contrastive examples illustrating the substantial variations within one event. Both video sequences are of the event "*marriage proposal*" in the TRECVID MED dataset. The first event took place in a classroom while in the second video a man proposed outdoor. People with different cultural background will also have very different marriage proposals, e.g., the western and eastern marriage proposals can be very different. Compared to concept (scene, action, object, etc.) analysis, event detection is more challenging in the following aspects:

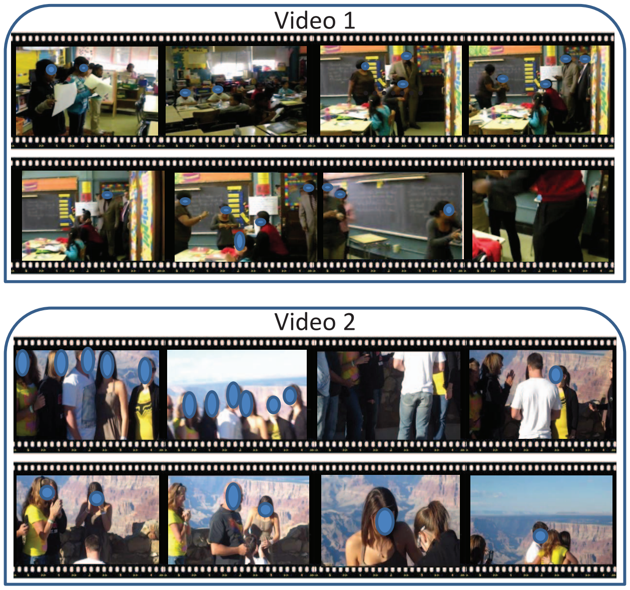


Figure 1. Two video sequences of the event "*marriage proposal*" in TRECVID MED dataset. An event may take place in different places with huge variations in terms of lighting, resolution, duration and so forth.

Firstly, an event is a higher level semantic abstraction of video sequences than a concept and consists of multiple concepts. For example, a "*marriage proposal*" event can be described by multiple objects (e.g., ring, faces), scene (e.g., in a restaurant), actions (e.g., talking, kneeling down) and acoustic concepts (e.g., music, cheering).

Secondly, a concept can be detected in a shorter video sequence or even in a single frame but an event is usually contained in a longer video clip. The object "*table*" can be detected in a single frame and the action "*jump*" only occurs in a very short video clip. In contrast, a video sequence of the event "*birthday party*" may last longer. If we see only a few frames showing some people chatting, we could not know if it is a “ *birthday party*” event or not.

Thirdly, different video sequences of a particular event may have dramatic variations. Taking “*giving directions to a location*” event as an example, it may take place in the street, inside a shopping mall or even in a car, where the visual features are very different. Also, web videos have huge variations. For example, they can be recorded either by a mobile camera with fast camera motion or by a professional video recorder fixed on a tripod.

While much progress has been made on visual concept recognition recently, the detection of complex event is still in its infancy. Most previous works of video concept/event detection are constrained to the detection of unusual activities (but not typical events) in small video dataset, e.g., abnormal activity or repetitive patterns detection in video sequence. Since 2012, limited studies focusing on complex event analysis of web videos have been reported. In [6], researchers proposed a graph based approach to analyze the relationship among different concepts such as action, scene, and object for complex event analysis. However, they only focused on event *recognition* while event *detection* is a more challenging task. Tamrakar et al. have experimentally compared seven visual features for complex event detection in web videos [14] and found that MoSIFT [3] is the most discriminative feature [14]. Ma *et al*. proposed to adapt the auxiliary knowledge from pre-labeled video dataset to facilitate event detection [9] where only 10 positive exemplars are available. The study in [12] has combined acoustic feature, texture feature and visual feature for event detection. Xu *et al*. have proposed an decision level fusion algorithm, which jointly considers threshold and smoothing factor to learn optimal weights of multiple features, for event detection [18]. In literature, Support Vector Machine (SVM) with χ^2 kernel has been shown to be an effective tool for event detection in research papers and TRECVID competition [20] [11] [9] [12] [14]. In [10], event detection and video attribute classification are integrated into a joint framework to leverage the mutual benefit.

Compared to concepts, an event is a higher level abstraction of a longer video clip. We should not directly apply the method proposed for concept recognition with weak supervision, *e.g*. one shot learning [4] or attribute based recognition [1], to our problem. Due to the semantic richness of an event in longer web videos, we may need more positive exemplars for training. For example, if all the positive exemplars of “*marriage proposal*” we have are indoor videos, the system probably may not be able to detect the second video in Figure 1 as “*marriage proposal*.” In addition, many frames in web videos are not semantically meaningful. As shown in Figure 1, only a small portion of the frames are directly related to marriage proposal.

Last year, TRECVID Multimedia Event Detection (MED) Track has launched a new task, aiming to detect complex event in web videos when only 10 positive and 10 related video exemplars are available. The premise is that it is a non-trivial task to collect a positive exemplar video which conveys the precise semantic of a particular event and excludes any irrelevant information. It is comparatively easier to collect a video exemplar which is related to a particular event, but does not necessarily contain all the essential elements of the event. A main problem confronted is that it remains unclear how to use the related exemplars, as they are neither positive nor negative. The related exemplars can be of any other event, e.g., both “*thesis proposal*” and “*people date*” are related to “*marriage proposal*”. Thus transfer/multi-task learning does not apply to this problem as it remains unclear how to set the labels of the related exemplars. Due to the difficulties, although NIST has provided related exemplars for event detection in TRECVID, none of the existing systems has ever used these data. In this paper, we aim to detect complex events using only 10 positive exemplars along with 10 related video exemplars for event detection. To the best of our knowledge, this paper is the first research attempt to automatically assess the relatedness of each related exemplar and utilize them adaptively, thereby resulting in more reliable event detection when the positive data are few.

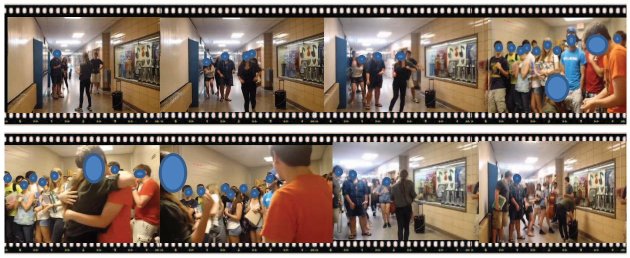


Figure 2. A video related to “*marriage proposal*.” A girl plays music, dances down a hallway in school, and asks a boy to prom.

2. Motivations and Problem Formulation

Detecting complex event using few positive exemplars is more challenging than the existing works which use more than 100 positive exemplars for training [12] [14]. Figure 2 shows some frames from a video clip marked as *related* to the event “*marriage proposal*” in TRECVID MED dataset. The video has several elements related to “*marriage proposal*,” *e.g*., a young man and a young lady doing something intimately, and people cheering. If we have sufficient positive exemplars for a particular event, including the related exemplars may not improve the performance. However, given that only few positive exemplars are available, it is crucial to make the utmost use of all the information.

Related exemplars are easier to obtain, but are much more difficult to use. The main problem is that the definition of “relatedness” is rather vague and subjective. There

Figure 3. A video related to "*marriage proposal*." A large crowd cheers after a boy asks his girlfriend to go to prom with him with a bouquet of flowers and a huge sign.

are not objective criteria to evaluate how close an exemplar is related to the target event. Simply assigning identical labels to different related exemplars does not make much sense as a related exemplar can be either closely or loosely related to the target event. Let us consider the case shown in Figure 3: a young man approaches to a young lady with a bouquet of flowers, kneels down in front of the lady, then they kiss and hug, and a large crowd cheers afterwards. The video looks pretty much like a "*marriage proposal*" event but it is not. The young man in that video actually asks his girlfriend to go to prom with him as opposed to proposing marriage. Compared to Figure 2, the video sequence shown in Figure 3 is more likely to be a "*marriage proposal*" event. It would be better to label the video shown in Figure 3 as closer to positive exemplar than that shown in Figure 2. Consequently, adaptively assigning soft labels to related exemplars by automatically assessing the relatedness turns to an important research challenge.

Next, we give our algorithm which is able to assign labels to related exemplars adaptively. Suppose there are n training videos $\{x_1, x_2, ...x_m, x_{m+1}, ..., x_n\}$, which are grouped into three classes, i.e., positive exemplars, null videos and related exemplars, and $x_{m+1}, ...x_n$ are related exemplars. Hereafter, a null video is a video sequence which can be any video sequence except for positive and related exemplars. There are two label sets $\tilde{Y}$ and Y used in our algorithm. The first label set $\tilde{Y}$ is the same as the traditional classifiers such as SVM and least square regression, which does not account for the related exemplars. More specifically, if x_i is a positive exemplar, the ith row of $\tilde{Y}$ is $[0, 1]$, otherwise $\tilde{Y}_i = [1, 0]$. To encode the information from related exemplars, we introduce a new label set Y to infer the soft labels of positive and related exemplars.

The key problem is to adaptively infer a soft label reflecting the positiveness for each of the related exemplars. Denote $S_i \geq 0$ as a positive variable. We use a vector $A \in \mathbb{R}^n$ to indicate if a training data is a positive or related exemplar. If x_i is a positive exemplar, $A_i = 1$; if x_i is a related exemplar, $A_i = -1$; if x_i is null video, $A_i = 0$. Recall that in $\tilde{Y}$ used in the traditional classifiers, the label of a positive exemplar is set as 1 for the second column. Therefore, given a related exemplar x_i, its adaptive soft label Y_i^a should be $Y_i^a = 1 - S_i$. To better differentiate related and positive exemplars, if x_i is a positive exemplar, its adaptive soft label Y_i^a is set to be $Y_i^a = 1 + S_i$. If x_i is a null video, then its label is the same as in $\tilde{Y}$. The intuition lying behind is that related exemplars have positive attributes but less positive than the true positive exemplars. For negative exemplars, their labels are -1, which is the same as $\tilde{Y}$ used in the traditional classifiers. Denote $X = [x_1, ..., x_n]$. The basic model to adaptively assess the positiveness of related exemplars is formulated as follows:

$$\min_{P,S,Y^a} \left\| X^T P - Y \right\|_F^2 + \Omega(P)$$
$$s.t. \quad Y = [\tilde{Y}_1, Y^a], \; Y^a = \tilde{Y}_2 + A \odot S, \, S \geq 0, \qquad (1)$$

where $\tilde{Y}_1$ is the first column of $\tilde{Y}$, P is the transformation matrix from video feature X to the soft labels Y, $\Omega(P)$ is a regularization term on P, and $\odot$ is the Hadamard product. Intuitively, the labels of related exemplars should be smaller than those of positive exemplars. Thus for a positive exemplar x_i, we add a positive variable S_i to $(\tilde{Y}_2)_i$ in (1). Likewise, for a related exemplar x_j, we subtract a positive variable S_j from $(Y_2)_j$. That is the reason why we impose non-negative constraint on $S \in \mathbb{R}^n$. As the adaptive label matrix Y^a is an optimization variable in (1), the model is able to adaptively utilize the related exemplars on a per-exemplar basis. When there are no related exemplars available, Y is the same as $\tilde{Y}$, and the algorithm will reduce to least square regression. Given a related exemplar x_i, S_i adaptively reflects its relatedness. If x_i is less related to the target event, a larger value S_i will be subtracted from $\tilde{Y}_2$.

As the number of null videos is much larger than those of positive and related exemplars, we further cluster the null videos into k clusters by K-means as preprocessing. In this way, the training exemplars are grouped into k negative sets and one positive set (including related exemplars). Denote $Y_r = [Y_r^1, Y_r^2, ..., Y_r^{k+1}] \in \{0, 1\}^{n \times (k+1)}$, whose first k columns correspond to the k negative clusters and the last column Y_r^{k+1} corresponds to positive and related samples. The same as (1), if x_i is a null video from the jth negative cluster, then $Y_r^{ij} = 1$; if x_i is a positive or related exemplar then $Y_r^{i(k+1)} = 1$. Further to the basic model shown in (1), we constrain that the transformation matrix P in (2) should have some common structure with the detector which is learnt based on positive and null exemplars only, as positive exemplars are more accurate than related ones. We propose to minimize the following objective function:

$$\begin{aligned} \min_{W,P,S,Y} & \left\| \tilde{X}^T W - \tilde{Y} \right\|_F^2 + \left\| X^T P - Y \right\|_F^2 \\ & + \alpha(\|W\|_F^2 + \|P\|_F^2) + \beta \left\| E \right\|_* \\ & s.t. \; Y = [Y_r^1, Y_r^2, ..., Y_r^k, Y_r^a], \\ & Y_r^a = Y_r^{k+1} + A \odot S, E = [W, P], \; S \geq 0, \end{aligned} \qquad (2)$$

where $\|\cdot\|_*$ is the trace norm of a matrix and $\tilde{X} = [x_1, ..., x_m]$. In (2), the trace norm minimization of $E = [W, P]$ is adopted to uncover the shared knowledge of W and P [19]. The minimization of (2) would allow the system to analyze the relationship between the positive and the related exemplars, and obtain the optimal adaptive label Y.

3. The Optimization Procedure

Let $D = \frac{1}{2}(EE^T)^{-\frac{1}{2}}$. We convert Eq (2) to:

$$\begin{aligned} &\min_{W,P,S} \left\|\tilde{X}^T W - \tilde{Y}\right\|_F^2 + \left\|X^T P - Y\right\|_F^2 \\ &+\alpha(\|W\|_F^2 + \|P\|_F^2) + \beta Tr(E^T D E) \\ &\qquad s.t.\ Y = [Y_r^1, Y_r^2, ..., Y_r^k, Y_r^a], \\ &\quad Y_r^a = Y_r^{k+1} + A \odot S, E = [W, P],\ S \geq 0, \end{aligned} \tag{3}$$

By setting the derivative of Eq (3) w.r.t. P to 0, we get:

$$P = (XX^T + \alpha I + \beta D)^{-1} XY. \tag{4}$$

Then we fix P and optimize W and S. By setting the derivative of Eq (3) w.r.t. W to 0, we get:

$$W = (\tilde{X}\tilde{X}^T + \alpha I + \beta D)^{-1} \tilde{X}\tilde{Y} \tag{5}$$

Optimizing S equals to the following problem:

$$\min_{S \geq 0} \left\|(X^T P)_{k+1} - (Y_r^{k+1} + A \odot S)\right\|_F^2, \tag{6}$$

where $(X^T P)_{k+1}$ denotes the $(k+1)$th column of $X^T P$. Let $(X^T P)_{k+1} - Y_r^{k+1} = M$, it becomes:

$$\min_{S \geq 0} \|M - A \odot S\|_F^2 \tag{7}$$

The optimal solution to (7) can be obtained by

$$S_{ij} = \max(M_{ij}/A_{ij}, 0). \tag{8}$$

Based on the discussion, we propose the algorithm shown in Algorithm 1 to optimize the objective problem and the convergence is guaranteed by Theorem 1.

Theorem 1. *Algorithm 1 monotonically decreases the objective function value of Eq (2) until convergence.*

Proof. According to Step 2 of Algorithm 1:

$$\begin{aligned} &\left\|\tilde{X}^T W_{t+1} - \tilde{Y}\right\|_F^2 + \left\|X^T P_{t+1} - Y_{t+1}\right\|_F^2 \\ &+\alpha(\|W_{t+1}\|_F^2 + \|P_{t+1}\|_F^2) + \beta Tr(E_{t+1}^T D_t E_{t+1}) \\ &\leq \left\|\tilde{X}^T W_t - \tilde{Y}\right\|_F^2 + \left\|X^T P_t - Y_t\right\|_F^2 \\ &+\alpha(\|W_t\|_F^2 + \|P_t\|_F^2) + \beta Tr(E_t^T D_t E_t) \end{aligned} \tag{9}$$

Algorithm 1: Adaptive Relatedness Analysis.

```
Input:
    $X \in \mathbb{R}^{d \times n}$, $\tilde{X} \in \mathbb{R}^{d \times m}$, $\tilde{Y} \in \mathbb{R}^{n \times 1}$,$Y_r \in \mathbb{R}^{n \times (k+1)}$;
    Parameters $\alpha$ and $\beta$.
Output:
    Optimized $W \in \mathbb{R}^{d \times 1}$, $P$ and $S$.
1: Set $t = 0$ and initialize $W$ and $P$ randomly;
2: repeat
    Compute $D_t$ as: $D_t = \frac{1}{2}(E_t E_t^T)^{-\frac{1}{2}}$ ;
    Update $P_t$ according to Eq (4);
    Update $W_t$ according to Eq (5) ;
    Compute $X^T P_t - Y_r = M_t$;
    Compute $S_t$ by $S_t^{ij} = \max(M_t^{ij}/A_{ij}, 0)$;
    $t = t + 1$.
   until Convergence;
```

Substituting $D_t = \frac{1}{2}(E_t E_t^T)^{-\frac{1}{2}}$ into Eq (9), it becomes:

$$\begin{aligned} &\left\|\tilde{X}^T W_{t+1} - \tilde{Y}\right\|_F^2 + \left\|X^T P_{t+1} - Y_{t+1}\right\|_F^2 \\ &+\alpha(\|W_{t+1}\|_F^2 + \|P_{t+1}\|_F^2) + \frac{\beta}{2} Tr\left(E_{t+1} E_{t+1}^T (E_t E_t^T)^{-\frac{1}{2}}\right) \\ &\leq \left\|\tilde{X}^T W_t - \tilde{Y}\right\|_F^2 + \left\|X^T P_t - Y_t\right\|_F^2 \\ &+\alpha(\|W_t\|_F^2 + \|P_t\|_F^2) + \frac{\beta}{2} Tr\left(E_t E_t^T (E_t E_t^T)^{-\frac{1}{2}}\right) \end{aligned} \tag{10}$$

According to *Lemma 1* in [19]:

$$\begin{aligned} &\frac{\beta}{2} Tr\left(E_{t+1} E_{t+1}^T (E_t E_t^T)^{-\frac{1}{2}}\right) - \beta Tr(E_{t+1} E_{t+1}^T)^{\frac{1}{2}} \\ &\qquad \geq \frac{\beta}{2} Tr\left(E_t E_t^T (E_t E_t^T)^{-\frac{1}{2}}\right) - \beta Tr(E_t E_t^T)^{\frac{1}{2}}. \end{aligned} \tag{11}$$

Subtracting Eq (11) from Eq (10), we have:

$$\begin{aligned} &\left\|\tilde{X}^T W_{t+1} - \tilde{Y}\right\|_F^2 + \left\|X^T P_{t+1} - Y_{t+1}\right\|_F^2 \\ &+\alpha(\|W_{t+1}\|_F^2 + \|P_{t+1}\|_F^2) + \beta \|E_{t+1}\|_* \\ &\leq \left\|\tilde{X}^T W_t - \tilde{Y}\right\|_F^2 + \left\|X^T P_t - Y_t\right\|_F^2 \\ &+\alpha(\|W_t\|_F^2 + \|P_t\|_F^2) + \beta \|E_t\|_* \end{aligned} \tag{12}$$

As the objective function value of (2) is lower bounded by zero, the proposed algorithm converges. □

4. Experiments

In this section, we conduct extensive experiments to test our algorithm using large scale real world dataset.

4.1. The Dataset

In 2011, NIST collected a large scale video dataset, namely MED 11 DEV-O collection, as the test bed for event detection. MED 11 DEV-O was collected from a variety of Internet hosting sites, which consists of over 32,000 testing videos and the total duration of DEV-O collection is about 1,200 hours. It is so far the largest publicly available video dataset with fine human labels, either in terms of total duration or number of videos. There are 10 events defined by NIST for TRECVID MED 11 evaluation. In 2012, NIST added another 10 new events to the evaluation. NIST provided about 2,000 positive exemplars of the 10 new events as MED 12 Develop collection. There are two types of event detection tasks defined by NIST. The first one is to detect complex event using about 150 positive exemplars. The other one is to detect events using only 10 positive exemplars and 10 related exemplars. We use the 10 positive exemplars and the related exemplars of each event identied by NIST for training. 1000 null videos from MED 11 develop dataset are used as negative exemplars.

In our experiment, we use all the testing data in MED 11 DEV-O collection for the 10 MED 11 events. Since the labels for TRECVID MED 12 testing collection are not released, we remove the 10 positive and related exemplars from MED 12 Develop collection and merge the remaining into MED 11 DEV-O collection as the testing data. Given a testing video x_t, the detection score is given by $(P^T x_t)_{k+1}$.

4.2. Experiment Setup

We use three motion features in our experiments, including Space-Time Interest Points (STIP) [8], Motion SIFT (MoSIFT) [3], and Dense Trajectories [15]. These features utilize different descriptors to capture the shape and temporal motion information of videos. In addition, we use three static features from key frames in the experiment, including Scale-Invariant Feature Transform (SIFT), Color SIFT (CSIFT) [13] and Transformed Color Histogram (TCH). We have a 32,768 dimension spatial BoWs for each feature as we used in [20]. We use Blacklight at Pittsburgh Supercomputing Center, a cluster which has 4096 cores and 32 TB RAM, to extract the visual features. Up to 1000 cores were simultaneously used to extract the 6 visual features. However, it is worth noting that our algorithm is much faster than feature extraction. Both training and testing are run on a desktop once the features are extracted.

Leveraging related exemplars for event detection is so far an unexploited area. To the best of our knowledge, there is no directly related algorithm to compare with. Support Vector Machine (SVM) and Kernel Regression (KR) are the most widely used classifiers in TRECVID MED 11 competition by the top ranked teams and recent research papers [12] [14] [20] [11] [9]. Therefore, we mainly compare our algorithm to them. To show the advantage of our algorithm in utilizing related exemplars, we report the results of SVM and KR using related exemplars as positive exemplars, which are denoted as SVM$_{RP}$ and KR$_{RP}$ hereafter. In addition, as the related exemplars may not be closely related to the target event, we also report the results of SVM and KR using related exemplars as negative exemplars, which are denoted as SVM$_{RN}$ and KR$_{RN}$. Moreover, we compare our algorithm to the multi-task feature learning algorithm MTFL proposed in [7], which considers positive, negative, and related prediction as three tasks and learns the shared representations among tasks in the same group. Since we have three tasks in total, we use 2 as the group number, which makes related videos share some information with both positive ones and negative ones. Average Precision (AP) and Mean AP (MAP) are used as the evaluation metrics.

Table 1. Experiment results on LFW dataset.

Method	AP(%)	Method	AP(%)	Method	AP(%)
Ours	**22.6**	KR	19.0	KR$_{\text{RP}}$	18.3
KR$_{\text{RN}}$	18.7	SVM	19.5	SVM$_{\text{RP}}$	19.0
SVM$_{\text{RN}}$	19.2	MTFL	19.6	–	–

The χ^2 kernel as described in [16] has been demonstrated the most effective kernel for event detection [12] [14] [20] [11]. Therefore, we use the χ^2 kernel defined below for SVM, SVM$_{RP}$, SVM$_{RN}$, KR, KR$_{RP}$ and KR$_{RN}$. Following [9] [10], we perform full-rank Kernel PCA (KPCA) with χ^2 kernel to get the kernelized representation of videos for our algorithm and MTFL. The parameters of all the algorithms are searched on the grid of $\{10^{-4}, 10^{-2}, 1, 10^2, 10^4\}$. After KPCA, K-means should be performed to have a more balanced input for each feature. k is empirically set to 10 in our experiment.

4.3. Experiment on A Toy Problem

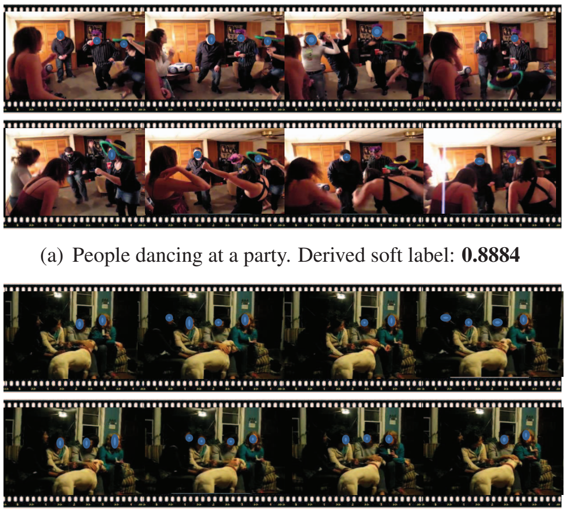
We use the LFW dataset [5] containing face images of 5749 individuals as a toy problem. The dataset is not equally distributed. For example, there are 530 images of George Walker Bush, but only 13 images of George Herbert Walker Bush. The 530 images of George W. Bush are divided into two subset: 10 as positive training exemplars and the remaining 520 data for testing. We then use the 13 face images of George H.W. Bush as related exemplars because of the father-child relationship. To test the robustness of the algorithm, we add some noises to six (about a half) of the face images of George H.W. Bush. We sample one image for each person as background images, resulting in 5746 background images. 1000 null images are used as negative training exemplars and the remaining 4746 images are used as null testing images. In this toy, pixel values are used as image features. Experiment results are reported in Table 1, which shows that our algorithm gains the best performance.

4.4. Experiment on Complex Event Detection

We first use some examples to show how the proposed algorithm adaptively assigns soft labels to related exemplars.

Table 2. Mean Average Precision (MAP) (%) of the 20 events. The proposed algorithm significantly outperforms all the other algorithms.

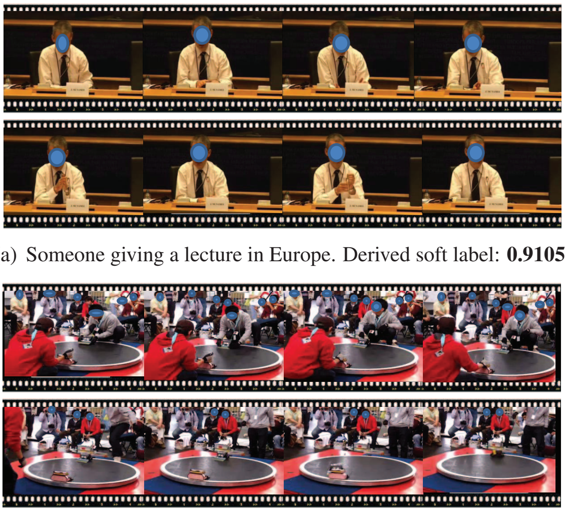
Feature	Our Algorithm	KR	KR_{RP}	KR_{RN}	SVM	SVM_{RP}	SVM_{RN}	MTFL
STIP	**4.9**	4.5	4.1	4.4	4.0	3.6	1.1	4.5
MoSIFT	**7.1**	6.7	6.3	6.6	6.2	5.4	1.0	6.7
Trajectory	**8.9**	8.3	8.0	8.2	7.8	7.1	1.9	8.1
SIFT	**6.8**	6.3	6.2	6.2	6.3	6.1	1.0	6.3
TCH	**4.3**	3.9	3.8	3.9	3.9	3.6	0.8	3.9
CSIFT	**5.6**	5.2	4.5	5.1	5.2	4.4	0.9	5.1
Average	**6.3**	5.8	5.5	5.7	5.6	5.0	1.1	5.8



(a) People dancing at a party. Derived soft label: **0.8884**

(b) People sitting in a house at night. Derived soft label: **0.4761**

Figure 4. The frames sampled from two video sequences marked as related exemplars to the event "*birthday party*" by NIST.



(a) Someone giving a lecture in Europe. Derived soft label: **0.9105**

(b) Two guys demonstrating robot fight. Derived soft label: **0.6581**

Figure 5. The frames sampled from two video sequences marked as related to the event "*town hall meeting*" by NIST.

In (2) the label matrix Y is an optimization variable as apposed to a fixed constant. Given a related exemplar, if it is closely related to the target event, the corresponding learned label should be larger than the one which is loosely related to the target event. Figure 4 shows the frames sampled from two video sequences marked as related exemplars of the event *"birthday party"* by NIST. The first video is not a "*birthday party*" event, but is about people singing and dancing in a party. It is pretty much related to "*birthday party*". In the second video, a few people sit and chat at night. It is also related to a "*birthday party*" event as there are several people in the video. One can see that the first video is more related to the target event. Correspondingly, the derived soft label of the first video is 0.8884 and the derived soft label of the second one is only 0.4761, both of which are quite consistent with human perception. Figure 5 shows another example, which includes the frames sampled from two related exemplars of the event "*town hall meeting*". The system learned soft label for the video in which someone gives a lecture in Europe is 0.9105. The soft label of the other video in which two guys demonstrate robot fight for audience is 0.6581. We can see that using the learned optimal label for training makes more sense than simply taking them as positive or negative exemplars. These examples also demonstrate that it is less reasonable to fix the labels of related exemplars as a smaller constant, *e.g*. 0.5.

Next, we quantitatively compare the proposed algorithm to the state of the art. Table 2 shows the MAP of all the 20 events using different visual features. Our algorithm outperforms KR by almost 9% relatively, indicating that it is beneficial to utilize related exemplars for event detection. As human assessment of relatedness is subjective, the selection of related exemplars is somewhat arbitrary. Some of the related exemplars should be regarded as positive exemplars but others are much less positive. As a result, we observe from Table 2 that KR outperforms both KR$_{RN}$ and KR$_{RP}$. Similarly, SVM also outperforms SVM$_{RN}$ and SVM$_{RP}$. This observation indicates that different related exemplars should be utilized adaptively. Using related exemplars as either positive or negative will degrade the overall performance for both SVM and KR. That could be also the reason why none of the existing event detection systems built in 2012 used related exemplars for event detection [20] [11] [9] [14], although NIST has provided them.

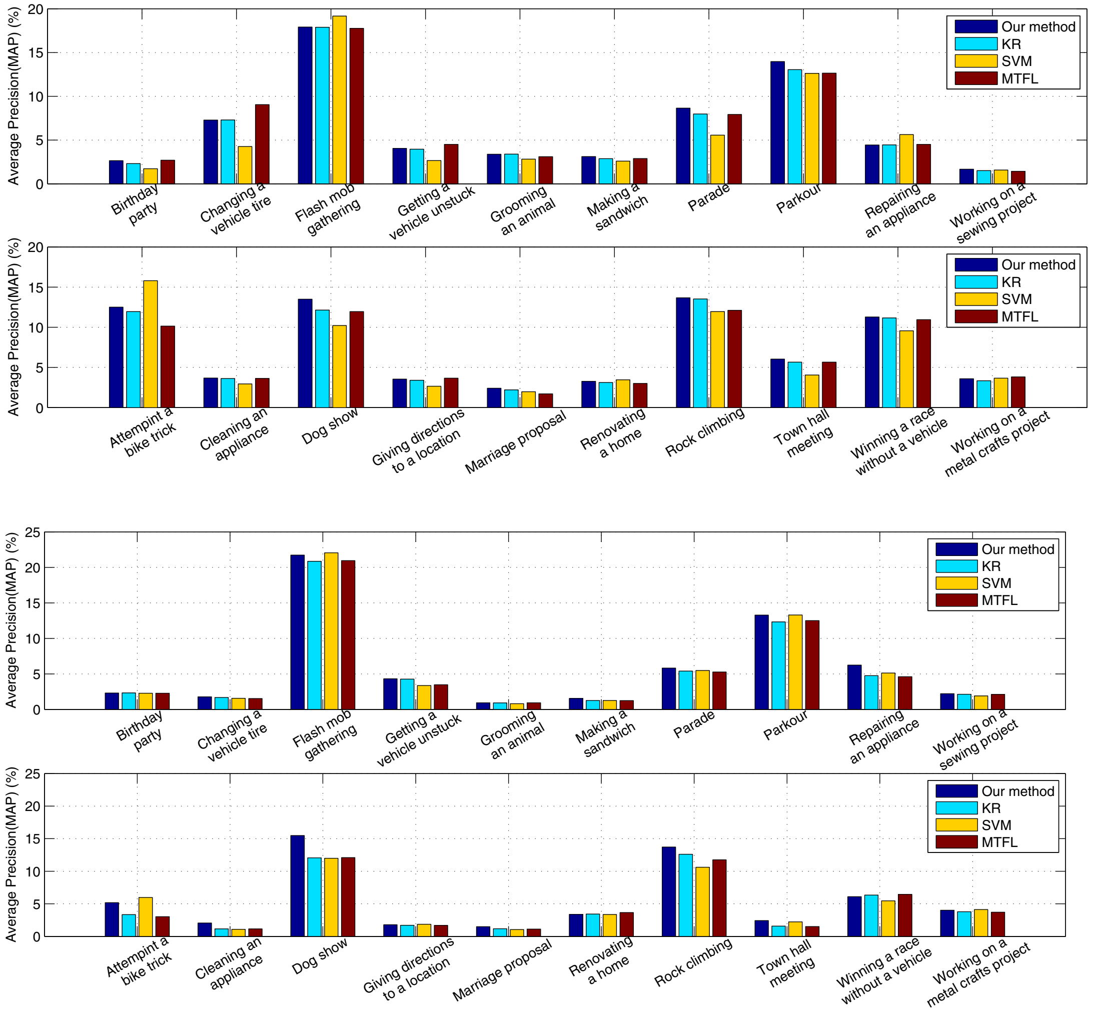


Figure 6. Performance comparison using MoSIFT (the upper two subfigures) and Color SIFT (the lower two subfigures) on MED dataset.

Next, taking MoSIFT and Color SIFT as showcases, we report the AP of each event. The upper two subfigures of Figure 6 show the event detection performance of using MoSIFT feature. The lower two subfigures of Figure 6 show the event detection performance of using Color SIFT feature. Looking into Figure 6, we observe that when using MoSIFT feature, our algorithm achieves the best or second best performance for 19 out of 20 events. When using Color SIFT feature, our algorithm achieves the best or second best performance for 18 out of 20 events. As SVM and Kernel regression models have been demonstrated very effective for complex event detection [12] [14] [20] [11] [9], this experiment demonstrates that our model not only gains the best MAP for all the events, the performance is also stable across multiple events.

4.5. The Limitations

Although the proposed algorithm gains promising performance in event detection, it still has some limitations. In our setting, the number of negative training data is much larger than that of positive ones. The major problem of the algorithm is that it may fail in relatedness analysis mainly due to the imbalanced numbers of positive and negative samples. As the number of positive examples are much smaller than that of negative examples, all the induced soft labels of related videos will be the same as negative labels. As a way to relieve the drawback, K-means clustering is performed to divide the negative samples into multiple classes. Figure 7 shows a failure example of our algorithm. If we look at the video, it is very similar to the event visually and should have a higher score. In the experiment the derived soft label is only 0.3136, which is not quite consistent with human supervisor's decision. In the future, we will study how to do a better job in dealing with imbalanced

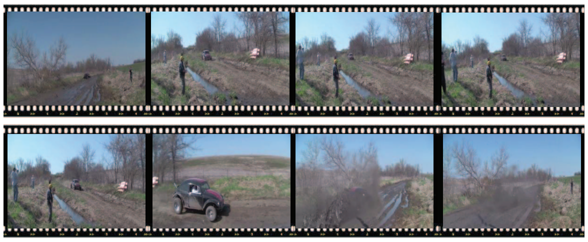


Figure 7. The frames sampled from a video sequence marked as related to the event "*Getting a vehicle unstuck*" by NIST.

training data for relatedness analysis.

5. Conclusions

We have focused on how to utilize related exemplars for event detection when the positive exemplars are few. The main challenge confronted is that the human labels of related exemplars are subjective. We propose to automatically learn the relatedness and assign soft labels to related exemplars adaptively. Extensive experiments indicate that 1) taking related exemplars either as positive or negative exemplars may degrade the performance; 2) our algorithm is able to effectively leverage the information from related exemplars by exploiting the relatedness of a video sequence. Future work will apply our model to interactive information retrieval where the users may not be able to get the exact search exemplars for relevance feedback.

6. Acknowledgements

All the features were extracted on Blacklight at Pittsburgh Supercomputing Center (PSC). We would like to thank PSC for providing the computing resource.

This work was partially supported by the ARC Project DE130101311, partially supported by the Singapore National Research Foundation under its International Research Centre @Singapore Funding Initiative and administered by the IDM Programme Office, and partially supported by 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.

References

[1] Z. Akata, F. Perronnin, Z. Harchaoui, C. Schmid, et al. Label-embedding for attribute-based classification. In *CVPR*, 2013.

[2] C. Cabrera, R. Sastre, J. Rodr, and S. Bas. Surfing the point clouds: Selective 3D spatial pyramids for category-level object recognition. *ICCV*, 2011.

[3] M. Chen and A. Hauptmann. Mosift: Reocgnizing human actions in surveillance videos. *Technical Report: CMU-CS-09-161*, 2009.

[4] L. Fei-Fei, R. Fergus, and P. Perona. One-shot learning of object categories. *TPAMI*, 28(4):594–611, 2006.

[5] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, October 2007.

[6] H. Izadinia and M. Shah. Recognizing complex events using large margin joint low-level event model. *ECCV*, 2012.

[7] Z. Kang, K. Grauman, and F. Sha. Learning with whom to share in multi-task feature learning. In *ICML*, 2011.

[8] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *CVPR*, 2008.

[9] Z. Ma, Y. Yang, Y. Cai, N. Sebe, and A. Hauptmann. Knowledge adaptation for ad hoc multimedia event detection with few examplars. *ACM Multimedia*, 2012.

[10] Z. Ma, Y. Yang, Z. Xu, S. Yan, N. Sebe, and A. Hauptmann. Complex event detection via multi-source video attributes. *CVPR*, 2013.

[11] P. Natarajan, P. Natarajan, A. Vazquez-Reina, S. Vitaladevuni, C. Andersen, R. Prasad, S.-F. Chang, I. Saleemi, M. Shah, Y. Ng, B. White, L. Davis, A. Gupta, and I. Haritaoglu. BBN VISER TRECVID 2012 Multimedia Event Detection and Multimedia Event Recounting Systems. In *TRECVID Workshop*, 2012.

[12] P. Natarajan, S. Wu, S. Vitaladevuni, U. Park, R. Prasad, and P. Natarajan. Multimodal feature fusion for robust event detection in web videos. *CVPR*, 2012.

[13] K. Sande, T. Gevers, and C. Snoek. Evaluating color descriptors for object and scene recognition. *TPAMI*, 32(9):1582–1596, 2010.

[14] A. Tamrakar, S. Ali, Q. Yu, J. Liu, O. Javed, A. Divakaran, H. Cheng, and H. Sawhney. Evaluation of low-level features and their combinations for complex event detection in open source video. *CVPR*, 2012.

[15] H. Wang, A. Klaser, C. Schmid, and C. Liu. Action recognition by dense trajectories. In *CVPR*, 2011.

[16] H. Wang, M. Ullah, A. Klaser, I. Laptev, C. Schmid, et al. Evaluation of local spatio-temporal features for action recognition. In *BMVC*, 2009.

[17] S. Wang, Y. Yang, Z. Ma, X. Li, C. Pang, and A. Hauptmann. Action recognition by exploring data distribution and feature correlation. *CVPR*, 2012.

[18] Z. Xu, Y. Yang, I. Tsang, N. Sebe, and A. Hauptmann. Feature weighting via optimal thresholding for video analysis. In *ICCV*, 2013.

[19] Y. Yang, Z. Ma, A. Hauptmann, and N. Sebe. Feature selection for multimedia analysis by sharing information among multiple tasks. *TMM*, 3(15):661–669, 2013.

[20] S. Yu, Z. Xu, D. Ding, et al. Informedia e-lamp @ TRECVID 2012, multimedia event detection and recounting. *TRECVID Workshop*, 2012.