Event Detection using Multi-Level Relevance Labels and Multiple Features

Zhongwen Xu[†] Ivor W. Tsang[‡] Yi Yang[†] Zhigang Ma[§] Alexander G. Hauptmann[§]
[†]ITEE, The University of Queensland, Australia
[‡]QCIS, University of Technology, Sydney, Australia
[§]School of Computer Science, Carnegie Mellon University, USA
{z.xu3,yi.yang}@uq.edu.au Ivor.tsang@uts.edu.au {kevinma,alex}@cs.cmu.edu

Abstract

We address the challenging problem of utilizing related exemplars for complex event detection while multiple features are available. Related exemplars share certain positive elements of the event, but have no uniform pattern due to the huge variance of relevance levels among different related exemplars. None of the existing multiple feature fusion methods can deal with the related exemplars. In this paper, we propose an algorithm which adaptively utilizes the related exemplars by cross-feature learning. Ordinal labels are used to represent the multiple relevance levels of the related videos. Label candidates of related exemplars are generated by exploring the possible relevance levels of each related exemplar via a cross-feature voting strategy. Maximum margin criterion is then applied in our framework to discriminate the positive and negative exemplars, as well as the related exemplars from different relevance levels. We test our algorithm using the large scale TRECVID 2011 dataset and it gains promising performance.

1. Introduction

A complex event is a higher level semantic abstraction of longer video clips than concepts such as actions, scenes or objects. For example, a "Birthday party" event may contain multiple concepts such as person, birthday cake, singing, cheering, blowing candles, *etc*. In addition, the videos depicting the same event usually have huge within event variations. A "Birthday party" may take place indoors (*e.g*. in a restaurant) or outdoors (*e.g*. in a park), and people may celebrate the birthday in different ways. They may sing a song, have a dinner, or play games. In contrast to actions that usually last for a few seconds and objects which can be detected in a single image, complex events generally have longer duration from several minutes to hours.

Even though it is more difficult, complex event detection has gradually attracted more research attention in recent years. Videos with long durations and large intra-class variances contain rich and complex information that the whole information of the videos cannot be captured by one single feature. Research papers and existing systems have demonstrated that combining multiple features is an effective method for event detection [25, 13, 16]. Some features are better for discriminating scenes, while others may be more sensitive to different actions. However, most systems combine these features in a simple way without considering correlations between different features for event detection [25, 13]. Intuitively, these features are correlated and complementary to each other. If we manage to uncover such shared information between different features, we will be able to leverage the mutual benefit of multiple features, which in turn will result in better exploitation of them. In light of this, we build up a multiple feature learning framework for complex event detection. Our framework is capable of handling different features jointly by appropriately mining their correlations, thus leading to a more robust event detector.

Figure 1. A video of "Man performs an oil change on a motorcycle", which is related to the event "Changing a vehicle tire".

Due to the complex attribute of an event, it is comparatively hard to find positive exemplars which exactly match the definition of the event. However, it is easier to find videos that match the definition partially, which is referred as related exemplars in this paper. As shown in Figure 1, a video described as "Man performs an oil change on a motorcycle" is marked as a related exemplar to the event "Changing a vehicle tire" by NIST. The action "change" and the object "motorcycle" are basic components of the "Chang-

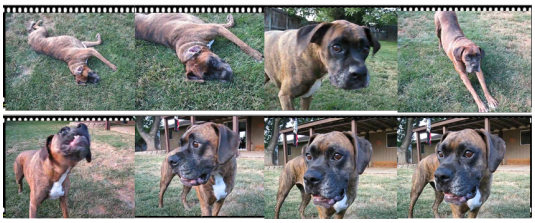

Figure 2. A video of "A dog lies in the grass", which is related to the event "Grooming an animal".

ing a vehicle tire" event, though the video has nothing to do with "tire". Figure 2 shows another example. A video depicted as "A dog lies in the grass" is considered as a related exemplar to the event "Grooming an animal" because it contains a dog, which belongs to the concept "animal".

Related exemplars share some common components with the target event, but at different levels. In Figure 3, we show four related videos to the event "Changing a vehicle tire". Some related videos, *i.e*. subfigures (c) and (d), are quite close to the positive exemplars, while others, *i.e*. subfigures (a) and (b), are just marginally related. Related exemplars are noisy but they essentially have useful information for discriminating an event. If we effectively utilize these related training exemplars, the detection performance would be boosted. While many papers and systems [25, 13, 16] have been proposed for complex event detection, only one has studied to use related exemplars for event detection [18]. However, as shown in Figure 3, the relevance levels of related exemplars vary dramatically. It is not reasonable to utilize them identically. The experiment in [18] also shows that using related videos as positive exemplars gains better performance for some events while for other events we should use related videos as negative exemplars. Even though cross validation could be potentially used, the algorithm proposed in [18] is not able to decide whether related videos should be used as positive or negative exemplars for a specific event, which severely limits the usage of the algorithm. In addition, [18] is not able to deal with multiple features.

Using all of the related videos as positive exemplars may increase false alarm as some of them do not necessarily have sufficient positive elements. Discarding related videos which are highly related to the target event, on the other hand, may lose useful information. Thus it is more reasonable to adaptively learn the relevance level of each related video and leverage the related videos of high relevance to infer a robust detector. Instead of directly using binary labels as in [13, 16, 25], we introduce ordinal labels to differentiate relevance levels of related videos. Specifically, if we use total R ($R \geq 3$) ordinal labels to denote the R relevance levels, we assign 1 as negative label, and R as positive label. The numbers between 1 and R correspond to related videos. A larger ordinal label indicates a higher relevance

(a) People driving on a road trip.

(b) Little kids washing cars.

(c) Two men working together to inflate a tire on a truck.

(d) A little boy trying to unscrew a bolt from a tire.

Figure 3. Related videos to the event "Changing a vehicle tire" provided by NIST. Different videos have different relevance levels to the event.

level. The ordinal labels close to 1 are the labels with low relevance to the event, and the ordinal labels close to R are the labels with high relevance.

There has been plenty of research focusing on multiple feature fusion [16, 15, 13], but none of the existing algorithms can tackle multiple relevance levels of the training data. To progress beyond the state of the art, we propose a cross-feature reasoning approach to generate a set of candidate labels for all related videos and then adaptively select an optimal ordinal label for each of them. After assigning one candidate label to each video, we enumerate possible combinations of all the related videos. We then learn an optimal weight to each label combination. In conjunction with a kernel matrix, each label combination can be used to train a model for event detection. Given multiple label combinations, we have multiple models. Then we formulate the label weighting problem in a multiple kernel learning fashion to obtain a unified event detector, where maximum margin criterion is applied to learn $(R-1)$ discriminative boundaries between each pair of consecutive ordinal labels. To make the results more robust, we propose to recursively update the label combinations. Once we get the unified event detector, we use it to predict the labels of related videos and update the label combinations, which are then used for another round of learning. The procedure is repeated until convergence and the final unified detector is used for event detection.

2. Related Work

In this section, we briefly review the related work on complex event detection and multiple feature learning.

2.1. Complex Event Detection

Complex event detection is intrinsically different from most event detection work that only handles simple events that can be characterized by a single shot or a few frames [5, 2, 7]. Most of these events are unusual events or sports events that last for short time and have small intra-class variations. In contrast, complex events focused in this paper are much more complicated, occur in much longer videos and have huge intra-class variations.

Different strategies, either from feature perspective or from classification perspective, have been developed to improve the detection accuracy. A method of using multi-channel shape-flow kernel descriptors is proposed for event detection in [12]. Yang *et al.* design an approach that discovers date-driven concepts from multi-modality signals, based on which a sparse video representation is learned for event detection [23]. Natarajan *et al.* propose to combine multiple features from different modalities to improve multimedia event detection [13]. Tamrakar *et al*. have evaluated the performance of several mainstream features for complex event detection [16]. Fisher vectors have been exploited instead of bag-of-word histograms to integrate the state-of-the-art low-level descriptors to represent the videos for complex event detection [14]. On the other hand, transfer learning has been exploited by Ma *et al.* for complex event detection when there are only a few positive examples [10]. Izadinia *et al*. consider that a complex event consists of some low-level events that can be treated as latent variables for learning a latent SVM based model [4]. Liu *et al*. have designed a local expert forest model for score fusion from multiple classifiers in complex event detection [9]. In [11], Ma *et al.* have proposed to use video attributes to boost event detection performance.

The progress made by the aforementioned work is encouraging and leads to more research efforts on complex event detection. But little work has made endeavor to topics on how to effectively utilize related exemplars for boosted performance due to the essential arduousness of the problem. [22] is one of the first attempt utilizing related exemplars for event detection. The algorithm porposed in [22] is desinged for only one single feature. In this paper, we dwell in the usage of related exemplars under multiple feature condition, which is a new and challenging topic.

2.2. Multiple Feature Learning

Current methods of combining multiple features work in a conditionally independent way. There are two major categories of multiple feature combination methods. One is early fusion, which combines the kernel matrices before the training process, and then puts the combined kernel matrix into the classifier. The other one is late fusion, which combines the predictive values after the training process. Some algorithms proposed in the machine learning community have good performance for different applications. But none of them has focused on our problem: how to effectively embed related exemplars learning into the multiple feature learning framework? To progress beyond the state of the art, we propose a novel multiple feature learning method that is particularly tailored to leverage related exemplars for complex event detection. Recently, Ye *et al*. have proposed a rank minimization based approach to fuse prediction scores for event detection [24]. However, the time complexity of the algorithm cubic w.r.t. the number of testing data, making it unsuitable for large scale datasets.

3. The Proposed Algorithm

In our framework, the relevance levels of the related videos are represented by R ordinal labels $(1, 2, \ldots, R)$. A larger ordinal label indicates higher relevance to the event. Label 1 is the negative exemplars, and label R is for positive exemplars. We have two major steps in our algorithm. Firstly, we learn the model with the maximum margin criterion between the consecutive relevance labels from a label candidates set. Secondly, we update the label candidate set from the prediction of cross-feature. These two steps are repeated until convergence.

3.1. Multi-Relevance Levels Learning

Assuming we have P different features extracted from the videos, we formulate the multiple feature learning problem with uncertain labels as follows:

$$\min_{f^p, \mathbf{y}^p \in \mathcal{Y}^p} \sum_{p=1}^{P} \left(\|f^p\|^2 + C \sum_{i=1}^{n} \ell(f^p, \mathbf{x}_i^p, \mathbf{y}_i^p) \right), \qquad (1)$$

where f^p is the classifier for the p-th feature, $1 \leq p \leq P$, $\mathcal{Y}^p$ is the label candidate set generated from the information of other features, C is the trade-off parameter to control the complexity of the model, and $\ell(f, x, y)$ is the loss function for classifier f, video x, and label y.

Without loss of generalization, we assume that $f^p(x) = (\mathbf{w}^p)^T \phi(\mathbf{x}^p)$, where $\phi(\cdot)$ maps the p-th feature into a Hilbert space. Then we apply the ℓ_1-hinge loss on (1). For the relevance labels, assuming R ordinal labels $(1, 2, \ldots, R)$ are used to represent the relevance levels, we introduce $\boldsymbol{\theta}^p$ $(\theta_r^p \leq \theta_{r+1}^p, 1 \leq r \leq R-1)$ to represent the boundaries between relevance labels r and $r+1$ for the p-th feature. The binary prediction between labels r and $r+1$ is formulated as:

$$f'^p = \mathrm{sgn}((\mathbf{w}^p)^T \phi(\mathbf{x}^p) - \theta_r^p). \qquad (2)$$

According to the binary prediction formulation of (2), we transform the relevance labels into $(R-1)$ binary labels for each level. For example, if R = 4, relevance label 1 is extended to [-1, -1, -1], while relevance label 3 is extended

to [1, 1, -1]. Following the definition above, we reformulate (1) for the p-th feature as:

$$\min_{y^p \in \mathcal{Y}^p} \left\{ \min_{\mathbf{w}^p, \boldsymbol{\theta}^p, \rho^p, \boldsymbol{\xi}} \frac{1}{2} \|\mathbf{w}^p\|_2^2 + \frac{1}{2} \|\boldsymbol{\theta}^p\|_2^2 - \rho^p + C \sum_{r=1}^{R} \sum_{i=1}^{n} \xi_i^r \right.$$
$$\text{s.t. } y_i^r((\mathbf{w}^p)^T \phi(\mathbf{x}_i^p) - \theta_r^p) \geq \rho^p - \xi_i^r, \xi_i^r \geq 0$$
$$\forall i = 1, \ldots, n, \forall r = 1, \ldots, R-1,$$
$$\left. \theta_r^p \leq \theta_{r+1}^p, \forall r = 1, \ldots, R-1 \right\} \tag{3}$$

According to [8], since the hinge loss is non-increasing, the constraints $\theta_r^p \leq \theta_{r+1}^p$ are implicitly fulfilled in the optimization. Next we optimize (3) for each feature. Since (3) has the same optimization formulation for each feature, we omit the superscript p in the following derivation.

Introducing $\alpha_i^r \geq 0$ and $\lambda_i^r \geq 0$ ($1 \leq i \leq n, 1 \leq r \leq R-1$) as the Lagrangian variables, we have the Lagrangian of the inner minimization problem of (3) as:

$$L = \frac{1}{2} \|\mathbf{w}\|_2^2 + \frac{1}{2} \|\boldsymbol{\theta}\|_2^2 - \rho + C \sum_{r=1}^{R-1} \sum_{i=1}^{n} \xi_i^r$$
$$- \sum_{i=1}^{n} \sum_{r=1}^{R-1} \alpha_i^r (y_i^r (\mathbf{w}^T \phi(\mathbf{x}_i) - \theta_r) - \rho + \xi_i^r) - \sum_{r=1}^{R-1} \sum_{i=1}^{n} \lambda_i^r \xi_i^r \tag{4}$$

Taking the derivatives w.r.t. $\mathbf{w}$, $\boldsymbol{\theta}$, α_i^r and λ_i^r to zero, and substituting the results back into (4), we get the dual form of the inner optimization subproblem of (3) as:

$$\max_{\alpha} -\frac{1}{2} \sum_{i,j=1}^{n} \sum_{r,r'=1}^{R} \alpha_i^r \alpha_j^{r'} y_i^r y_j^{r'} K(x_i^r, x_j^{r'}), \tag{5}$$

where $K(\mathbf{x}_i^r, \mathbf{x}_j^{r'}) = \phi(\mathbf{x}_i^r)^T \phi(\mathbf{x}_j^{r'}) + \delta(r = r')$. Note that function $\delta(\cdot) = 1$ when the condition inside it holds, otherwise $\delta(\cdot) = 0$. To transform (5) into matrix form, we make $\boldsymbol{\alpha} = [\alpha_1^1, \ldots, \alpha_1^{R-1}, \ldots, \alpha_n^1, \ldots, \alpha_n^{R-1}]^T$, and $\hat{\mathbf{y}} = [y_1^1, \ldots, y_1^{R-1}, \ldots, y_n^1, \ldots, y_n^{R-1}]^T$, then (5) can be simplified to:

$$\min_{\hat{\mathbf{y}} \in \mathcal{Y}} \left\{ \max_{\boldsymbol{\alpha} \in \mathcal{A}} -\frac{1}{2} \boldsymbol{\alpha}^T (\mathbf{K} \odot \hat{\mathbf{y}} \hat{\mathbf{y}}^T) \boldsymbol{\alpha} \right\}, \tag{6}$$

where $\mathcal{A}$ is the feasible set of $\boldsymbol{\alpha}$, and $\mathcal{A} = \{\boldsymbol{\alpha} | \sum_{i=1}^{n} \sum_{r=1}^{R-1} \alpha_i^r = 1, 0 \leq \alpha_i^r \leq C, 1 \leq i \leq n, 1 \leq r \leq R-1\}$. Furthermore, according to the minimax inequality proposed in [6], (6) is lower-bounded by

$$\max_{\boldsymbol{\alpha} \in \mathcal{A}} \left\{ \min_{\hat{\mathbf{y}} \in \mathcal{Y}} -\frac{1}{2} \boldsymbol{\alpha}^T (\mathbf{K} \odot \hat{\mathbf{y}} \hat{\mathbf{y}}^T) \boldsymbol{\alpha} \right\}, \tag{7}$$

which is obtained by noticing sets $\mathcal{A}$ and $\mathcal{Y}$ are compact sets, and swapping $\min_{\hat{\mathbf{y}} \in \mathcal{Y}}$ and $\max_{\boldsymbol{\alpha} \in \mathcal{A}}$ makes (6) an upper bound of (7).

To derive the dual form of the inner optimization problem of (7), we reformulate (7) as:

$$\max_{\boldsymbol{\alpha} \in \mathcal{A}} \left\{ \max_{\Delta} -\Delta, \right.$$
$$\left. \text{s.t. } \Delta \geq \frac{1}{2} \boldsymbol{\alpha}^T (\mathbf{K} \odot \hat{\mathbf{y}}_m \hat{\mathbf{y}}_m^T) \boldsymbol{\alpha}, \forall \hat{\mathbf{y}}_m \in \mathcal{Y} \right\}. \tag{8}$$

By introducing Lagrangian variables $d_m \geq 0$ on the inequality constraint in (8), the above problem can be relaxed into:

$$\max_{\boldsymbol{\alpha} \in \mathcal{A}} \left\{ \min_{\mathbf{d} \in \mathcal{D}} -\frac{1}{2} \boldsymbol{\alpha}^T \left(\sum_{m: \hat{\mathbf{y}}_m \in \mathcal{Y}} d_m \mathbf{K} \odot \hat{\mathbf{y}}_m \hat{\mathbf{y}}_m^T \right) \boldsymbol{\alpha} \right\}, \tag{9}$$

where $\mathcal{D}$ is the feasible set of $\mathbf{d}$, and $\mathcal{D} = \{\mathbf{d} | \sum_{m: \hat{\mathbf{y}}_m \in \mathcal{Y}} d_m = 1, d_m \geq 0, \forall m : \hat{\mathbf{y}}_m \in \mathcal{Y}\}$.

Swapping $\max_{\alpha \in \mathcal{A}}$ and $\min_{d \in \mathcal{D}}$, (9) is equivalent to:

$$\min_{\mathbf{d} \in \mathcal{D}} \left\{ \max_{\boldsymbol{\alpha} \in \mathcal{A}} -\frac{1}{2} \boldsymbol{\alpha}^T \left(\sum_{m: \hat{\mathbf{y}}_m \in \mathcal{Y}} d_m \mathbf{K} \odot \hat{\mathbf{y}}_m \hat{\mathbf{y}}_m^T \right) \boldsymbol{\alpha} \right\}. \tag{10}$$

From (10) we can see that if we regard the matrices $\mathbf{K} \odot \hat{\mathbf{y}}_m \hat{\mathbf{y}}_m^T$, $m : \hat{\mathbf{y}}_m \in \mathcal{Y}$ as the base kernels in the Multiple Kernel Learning (MKL) problem, where the optimization leads to an optimal convex combination of $|\mathcal{Y}|$ base kernels.

3.2. Updating Label Candidates via Cross-Feature Voting

Due to the imbalanced data condition, we cannot get the precise threshold of deciding which predictive value should be labeled as $+1$ or -1. Instead of setting only one threshold (typically the threshold is 0) to discretize the predictive values into label sets, we utilize the predictive values from other features to incorporate their information, and then different thresholds are set on the predictive values to obtain the label candidate set. The intuition behind setting different thresholds is that under the imbalanced data condition, the predictive values are severely biased. So we need to set different thresholds, and then generate larger label candidate sets. Then our model can select good label candidates.

Regarding the source of predictive values, one may use predictive values only from the classifier trained in the last iteration. However, bad label candidates in the previous iterations may propagate into the following label candidates, and degrade the performance of the prediction. Especially, the label candidates may not be accurate during the first few iterations. Thus, except for the label candidates generated from the last iteration, we keep the ones from different iterations. We utilize the label candidates during the whole iteration process to make the generated label candidates more robust.

We propose a cross-feature voting approach to update the label candidates, in which the candidate labels of the p-th feature are voted by other features. The proposed approach is formulated as follows:

$$\mathcal{Y}_{t+1}^{p} = \left(\bigcup_{p' \neq p}^{P} \mathcal{Q}_{t}^{p'} \right) \bigcup \mathcal{Y}_{t}^{p}, \tag{11}$$

where $\mathcal{Q}_t^{p'}$ is obtained from projecting the predictive values in the t-th iteration of the p'-th feature into label candidates by setting different thresholds. In our algorithm, $\mathcal{Q}_t^{p'}$ is introduced to control the proportion of exemplars with high relevance. The proposed algorithm is summarized in Algorithm 1.

Algorithm 1: Multi-Relevance Level Learning with Multiple Features

Input: Video data from P different features: $\{\mathbf{x}^p\}$, initial labels under each feature: $\{\mathbf{y}_0^p\}$
Output: $\boldsymbol{\alpha}^p, \mathbf{d}^p, \mathcal{Y}^p$

```
Initialize Y^p = {y_0^p}, 1 ≤ p ≤ P;
m ← 1;
repeat
    for p ← 1 to P do
        Solve α^p and d^p based on Y^p according to (10);
        Obtain predictive values z^p on data points x^p;
        Set different thresholds on z^p and get the label candidates set Q^p which satisfies the constraint;
    end
    for p ← 1 to P do
        Y^p ← Y^p ∪ Q^{p'};
    end
    m ← m + 1;
until stop criterion;
```

3.3. Stop Criterion and Convergence

Note that (10) aims to minimize the objective function w.r.t. $\mathbf{d}$ and $\boldsymbol{\alpha}$, and the label candidate set $\mathcal{Y}$ is expanded in each new iteration. The objective value of the last iteration thus equals to setting $d_{m'}$ to zeros, for all new generated label candidates $\mathbf{y}_{m'}$. Subsequently, the objective function in the $(t+1)$-th iteration is smaller than that in the t-th iteration, which means that the objective function decreases monotonously with the iterations. When the change between objective value of the t-th iteration and that of the $(t+1)$-th iteration becomes a relatively small value, the algorithm can be regarded as converged. In our experiments, the proposed algorithm often converges within 10 iterations.

3.4. Time Complexity

The major cost is to solve the MKL problem in (10), where the complexity depends on the kernel type used in $\mathbf{K}$. If $\mathbf{K}$ is from a non-linear kernel, the MKL problem requires approximately $O(L((R-1)n)^{2.3})$, where L is the number of iterations inside the MKL and $O(((R-1)n)^{2.3})$ is the empirical complexity of SVM training complexity. Assuming that Algorithm 1 converges after T iterations, the total time complexity of the model is $O(TPL((R-1)n)^{2.3})$. However, when $\mathbf{K}$ is from a linear kernel, we can apply LIBLINEAR solver to get the solution of MKL, which has the time complexity of $O((R-1)n)$ instead of $O(((R-1)n)^{2.3})$. Therefore the total time complexity becomes $O(TPL(R-1)n)$ if LIBLINEAR is embedded to optimize the MKL problem, which is very efficient.

4. Experiments

4.1. The Dataset

In 2011, NIST collected a large number of videos from Internet hosting sites such as Youtube. The dataset (namely TRECVID MED) consists of more than 32,000 testing videos. The total duration is 1,200 hours which is about 800 GB in size. The events in the testing set of TRECVID MED 2011 are "Birthday party (BP)", "Changing a vehicle tire (CaVT)", "Flash mob gathering (FMG)", "Getting a vehicle unstuck (GaVU)", "Grooming an animal (GaA)", "Making a sandwich (MaS)", "Parade (PR)", "Parkour (PK)", "Repairing an appliance (RaA)", and "Working on a sewing project (WaSP)". The numbers of positive exemplars provided by NIST for each event vary from 100 to 200, and the number of related exemplars is around 150.

In the experiments, we use three visual features: Dense Trajectories [20], MoSIFT [1], and Color SIFT (CSIFT) [19], which have been shown to be among the best visual features in the TRECVID MED competition. We extract these three features from videos, then we generate the visual vocabulary with a size of 4,096 for each descriptor. The videos are mapped into 4,096 dimensional Bag-of-Words (BoWs). We apply 1x1, 2x2, and 3x1 spatial grids to generate the spatial-BoWs. Thus we have 32,768 dimensional spatial representation for each feature. According to [25, 21], χ^2-kernel is the most effective kernel for video analysis. So we apply KPCA on χ^2-kernel in the preprocessing stage of our proposed algorithm.

4.2. Compared Algorithms

According to the reports of top ranked teams in TRECVID competition and recent research papers on event detection [25, 13, 12], Support Vector Machine (SVM) and Kernel Regression (KR) are the most reliable algorithms for event detection. To illustrate the different results from different ways of utilizing related exemplars, we conduct fol-

lowing experiments. Firstly, we regard all of the related exemplars as with positive labels, train the event detectors, and report the results. We denote the experiments as SVM_{POS} and KR_{POS} respectively. Secondly, we treat all the related exemplars as negative exemplars, and these results are denoted as SVM_{NEG} and KR_{NEG} respectively. We apply the χ^2-kernel on SVM and KR, which is consistent with [25]. Regarding the algorithms of combining multiple features, we apply early fusion and late fusion on these features. For early fusion methods, we apply average early fusion. For late fusion methods, we use average late fusion and LP-Boost fusion [3].

In our proposed algorithm, we fix the parameter C as 1, utilize $R = 4$ for the relevance levels, and then we impose the constraint that the related videos with high relevance are not more than 20% of the total number of related videos. We use two evaluation metrics for comparison. One is Average Precision (AP) and the other is Pmiss, which is used in the official evaluation of TRECVID MED.

4.3. Experiment Results on Single Feature

Firstly, noting that we can obtain the predictive values for each feature in our model, we show the detection results of using one single feature. To observe the effectiveness of utilizing information from related labels, we compare to the results of SVM, SVM_{POS}, SVM_{NEG}, KR, KR_{POS}, and KR_{NEG}, with one single feature. Table 1 shows the detection performance (measured in Mean Average Precision over 10 events) of Dense Trajectories, MoSIFT, and CSIFT under different models. From Table 1 we can see that our algorithm outperforms other methods significantly with the three evaluated features respectively. Figure 4 shows the performance comparison over different models on the best single feature Dense Trajectories. We can observe consistent advantage of our method over other state-of-the-art models for various events. Specifically, our algorithm achieves the best performance among 8 of 10 events in terms of Pmiss and AP, except events MaS and RaA.

We observe from the results that utilizing the related exemplars improperly may degrade the performance. For example, SVM_{POS} gets worse results than SVM with all the three single features. Similar situation happens for KR: KR_{POS} gets worse performance than KR. These results indicate that the related exemplars cannot be regarded as positive directly. The relatedness defined in semantic level from human sense is unnecessarily very close to the positive exemplars. Hence, giving related exemplars with positive labels or high confidence is harmful to the performance. On the other aspect, we can see that SVM_{NEG} has similar performance to SVM, and KR_{NEG} has similar performance to KR. This may be interpreted as enlarging the space of negative exemplars does not make too many differences to the performance.

In our model, the ordinal labels give the power of discriminating "low relevance" and "high relevance" to the event detectors, which makes the model more flexible and more robust. Our model keeps the maximum margin criterion between the successive ordinal labels and updates the label candidates by setting different thresholds to the predictive values. In this way, our model has two advantages: the first advantage is that we can utilize the related exemplars in a more flexible and more reliable way, and the second one is that we can boost the performance of the single feature by utilizing the information from other features. The experimental results demonstrate the effectiveness of utilizing the related videos in multiple feature scenario.

4.4. Experiment Results on Multiple Features

In the single feature comparison experiments, using related exemplars as positive or negative does not bring performance improvement for SVM and KR. In this experiment, we only report the combination methods based on the prediction results without the related exemplars due to the space limit. We compare to the results from average late fusion of SVM (SVM_{late}), average late fusion of KR (KR_{late}), average early fusion of SVM ($\text{SVM}_{\text{early}}$), LPBoost fusion of SVM (SVM_{LP}), and LPBoost fusion of KR (KR_{LP}).

Detection performance measured in average Pmiss and mean Average Precision over 10 events are shown in Table 2, from which we can see that out algorithm achieves the best performance over other state-of-the-art multiple feature combination methods. To show the performance comparison to the results from recently published paper, we quote the results of the same settings from [17] directly. Tang *et al*. obtain mean AP of 0.2178 among the 10 events, and our algorithm achieves mean AP of 0.2507. It is noticed that Tang *et al*. combine 13 types of image features and 2 types of video features in MED dataset, so the number of features is several times of ours.

Figure 5 illustrates the detailed performance comparison for each event individually. We can see that our model obtains the best performance for various events. We achieve the smallest Pmiss for 7 out of 10 events, and the highest AP for 7 out of 10 events.

5. Conclusion

In this paper, we have focused on a challenging problem of utilizing related exemplars in the multiple feature scenario. We treat the relevance levels as ordinal labels and use the maximum margin criterion to discriminate the related videos between with low relevance and with high relevance. We formulate the problem as a label candidate selecting problem, which generates many label candidates and then selects the most appropriate ones in the learning process. The label candidates are adopted from the predictions of multiple features, merging newly generated candidates

Table 1. Detection performance comparison on single feature, Mean AP over total 10 events is reported.

Feature	Our algorithm	SVM	SVM_{POS}	SVM_{NEG}	KR	KR_{POS}	KR_{NEG}
Dense Trajectories	**0.2213**	0.1957	0.1916	0.1944	0.1967	0.1912	0.1961
MoSIFT	**0.1311**	0.1229	0.1182	0.1236	0.1224	0.1182	0.1234
CSIFT	**0.1625**	0.1425	0.1387	0.1452	0.1417	0.1377	0.1435

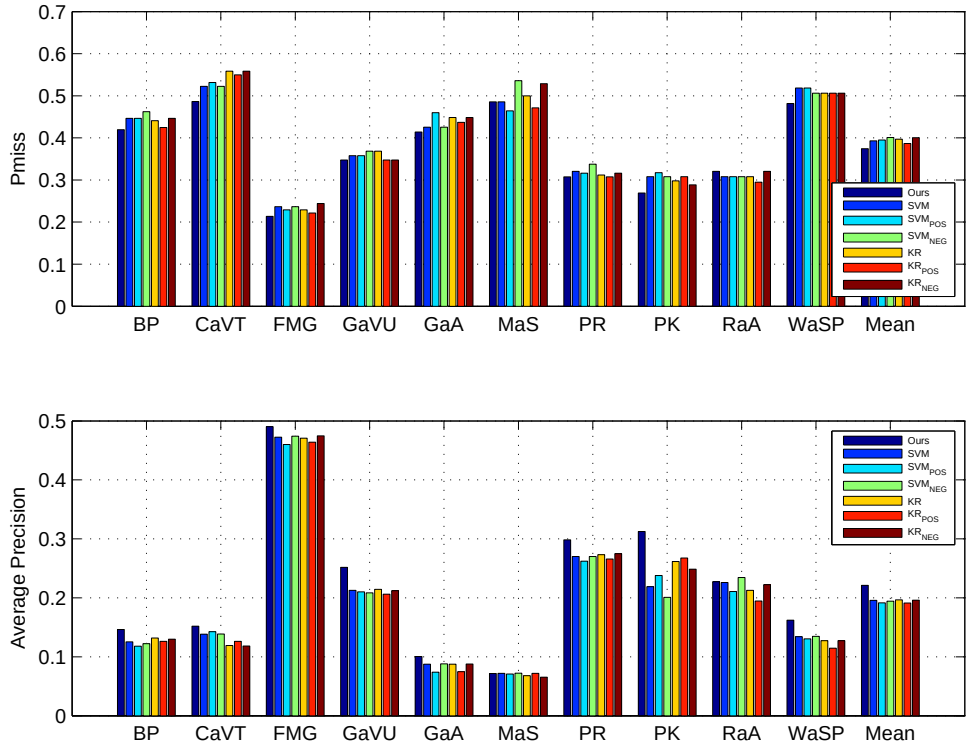

Figure 4. Performance comparison of different algorithms on the single feature Dense Trajectories. Full event names of the 10 events can be referred in Section 4.1. **LOWER** Pmiss values indicate **BETTER** performance. **HIGHER** AP values indicate **BETTER** performance.

in the iterations and setting different thresholds in the label prediction. Our proposed framework embeds the relevance levels learning problem into the multiple feature condition, which effectively utilizes the information contained in the related exemplars. Extensive experiments on TRECVID MED 11 dataset show the effectiveness of our framework, in which our proposed method outperforms other state-of-the-art methods significantly.

6. Acknowledgements

This work was partially supported by the ARC DECRA and Future Fellowship, partially supported by the Open Project Program of the State Key Lab of CAD&CG (Grant No. A1402), Zhejiang University, and partially supported by 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] M. Chen and A. Hauptmann. Mosift: Recognizing human actions in surveillance videos. In *CMU-CS-09-161*, 2009.

[2] Y. Cong, J. Yuan, and J. Liu. Sparse reconstruction cost for abnormal event detection. In *CVPR*, 2011.

[3] P. Gehler and S. Nowozin. On feature combination for multiclass object classification. In *CVPR*. IEEE, 2009.

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

[5] Y. Ke, R. Sukthankar, and M. Hebert. Efficient visual event detection using volumetric features. In *ICCV*, 2005.

[6] S. Kim and S. Boyd. A minimax theorem with applications to machine learning, signal processing, and finance. *SIAM Journal on Optimization*, 19(3):1344–1367, 2008.

[7] J. Kwon and K. M. Lee. A unified framework for event summarization and rare event detection. In *CVPR*, 2012.

Table 2. Detection performance comparison on three features, average Pmiss and average AP over total 10 events are reported. **LOWER** Pmiss values indicate **BETTER** performance. **HIGHER** AP values indicate **BETTER** performance.

Evaluation Metric	Our algorithm	SVM_{late}	KR_{late}	SVM_{early}	KR_{early}	SVM_{LP}	KR_{LP}
Pmiss	**0.3520**	0.3677	0.3704	0.3730	0.3711	0.3673	0.3634
AP	**0.2507**	0.2213	0.2235	0.2325	0.2268	0.2237	0.2225

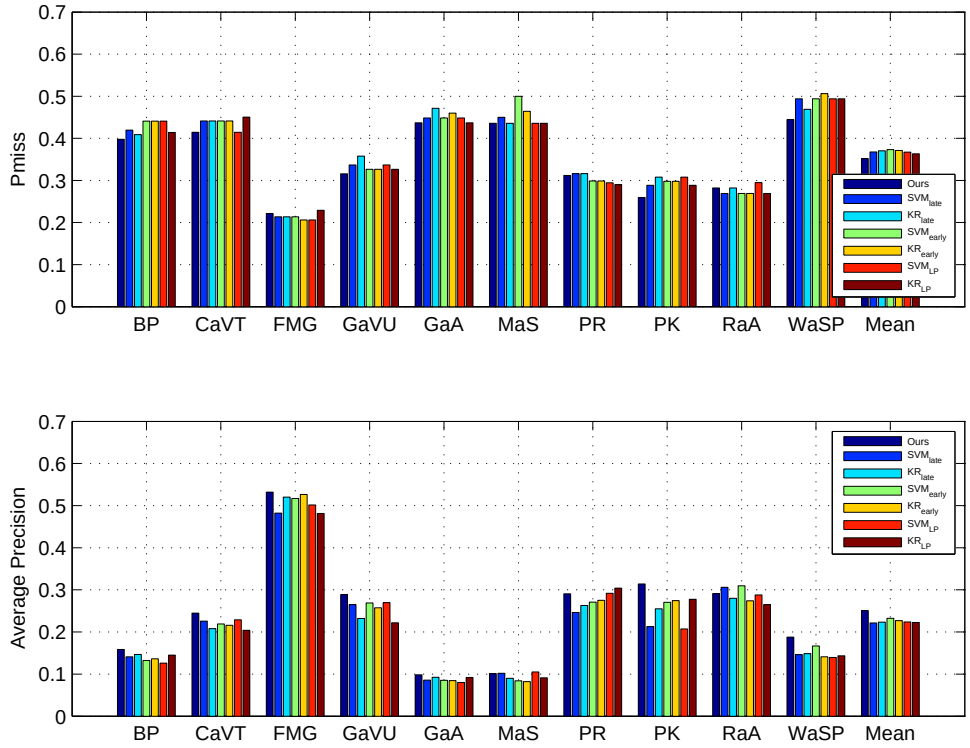

Figure 5. Performance comparison of different algorithms on three features. Full event names of the 10 events can be referred in Section 4.1. **LOWER** Pmiss values indicate **BETTER** performance. **HIGHER** AP values indicate **BETTER** performance.

[8] L. Li and H.-T. Lin. Ordinal regression by extended binary classification. In *NIPS*, 2006.

[9] J. Liu, S. McCloskey, and Y. Liu. Local expert forest of score fusion for video event classification. In *ECCV*, 2012.

[10] 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 Multimedia*, 2012.

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

[12] P. Natarajan, S. Wu, S. N. P. Vitaladevuni, X. Zhuang, U. Park, R. Prasad, and P. Natarajan. Multi-channel shape-flow kernel descriptors for robust video event detection and retrieval. In *ECCV*, 2012.

[13] P. Natarajan, S. Wu, S. N. P. Vitaladevuni, X. Zhuang, S. Tsakalidis, U. Park, R. Prasad, and P. Natarajan. Multimodal feature fusion for robust event detection in web videos. In *CVPR*, 2012.

[14] D. Oneata, J. Verbeek, and C. Schmid. Action and event recognition with fisher vectors on a compact feature set. In *ICCV*, 2013.

[15] A. Rakotomamonjy, F. Bach, S. Canu, and Y. Grandvalet. SimpleMKL. *JMLR*, 9:2491–2521, 2008.

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

[17] K. Tang, B. Yao, F.-F. Li, and D. Koller. Combining the right features for complex event recognition. In *ICCV*, 2013.

[18] C. Tzelepis, N. Gkalelis, V. Mezaris, and I. Kompatsiaris. Improving event detection using related videos and relevance degree support vector machines. In *Multimedia*. ACM, 2013.

[19] K. E. A. van de Sande, T. Gevers, and C. G. M. Snoek. Evaluating color descriptors for object and scene recognition. *TPAMI*, 32(9):1582–1596, 2010.

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

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

[22] Y. Yang, Z. Ma, Z. Xu, S. Yan, and A. G. Hauptmann. How related exemplars help complex event detection in web videos? In *ICCV*, 2013.

[23] Y. Yang and M. Shah. Complex events detection using data-driven concepts. In *ECCV*, 2012.

[24] G. Ye, D. Liu, I.-H. Jhuo, and S.-F. Chang. Robust late fusion with rank minimization. In *CVPR*, 2012.

[25] S. Yu, Z. Xu, D. Ding, W. Sze, F. Vicente, Z. Lan, Y. Cai, S. Rawat, P. Schulam, N. Markandaiah, et al. Informedia e-lamp@ trecvid2012: Multimedia event detection and recounting med and mer. In *NIST TRECVID Workshop*, 2012.