Complex Event Detection via Multi-Source Video Attributes

Zhigang Ma[†♮] Yi Yang[‡] Zhongwen Xu[‡§] Shuicheng Yan[♮] Nicu Sebe[†] Alexander G. Hauptmann[‡]
[†]Department of Information Engineering and Computer Science, University of Trento, Italy
[‡]School of Computer Science, Carnegie Mellon University, USA
[§]College of Computer Science, Zhejiang University, China
[♮]Department of Electrical and Computer Engineering, National University of Singapore, Singapore
{ma,sebe}@disi.unitn.it {yiyang,zhongwen,alex}@cs.cmu.edu eleyans@nus.edu.sg

Abstract

Complex events essentially include human, scenes, objects and actions that can be summarized by visual attributes, so leveraging relevant attributes properly could be helpful for event detection. Many works have exploited attributes at image level for various applications. However, attributes at image level are possibly insufficient for complex event detection in videos due to their limited capability in characterizing the dynamic properties of video data. Hence, we propose to leverage attributes at video level (named as video attributes in this work), i.e.*, the semantic labels of external videos are used as attributes. Compared to complex event videos, these external videos contain simple contents such as objects, scenes and actions which are the basic elements of complex events. Specifically, building upon a correlation vector which correlates the attributes and the complex event, we incorporate video attributes latently as extra informative cues into the event detector learnt from complex event videos. Extensive experiments on a real-world large-scale dataset validate the efficacy of the proposed approach.*

1. Introduction

In this paper, we focus on the event detection of large-scale real-world videos [2, 3]. An "event" refers to an observable occurrence that interests users and is found in specific scenes and is characterized by the subjects and objects involved [15]. In the past, detection of events that are simple, well-defined and describable by a short video sequence, *e.g.*, *hand shaking*, has been widely studied. In the real world, however, users are more interested in videos depicting complex events such as *celebrating the New Year*. Complex event detection is very challenging as these events usually contain many people and/or objects, various human actions, multiple scenes; have significant

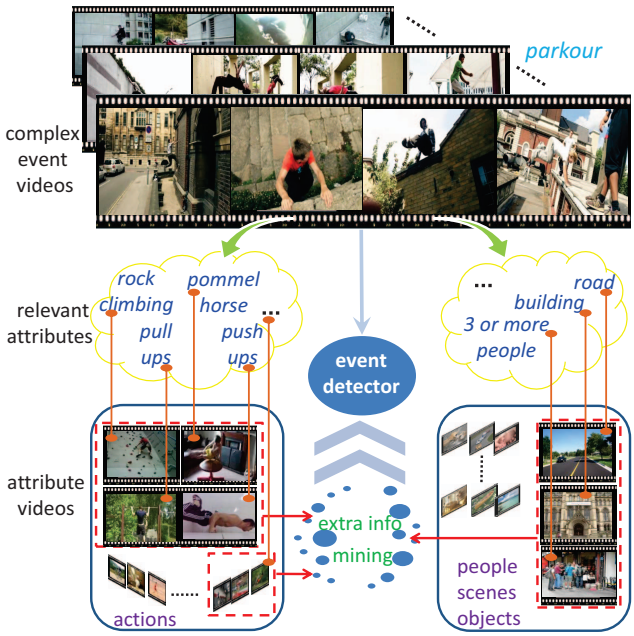

Figure 1. The illustration of our approach for complex event detection with video attributes.

intra-class variations; and take place in much longer video clips [2, 3, 15, 16]. Despite the arduousness, the practical significance of complex event detection has drawn increasing interest from researchers [23, 10, 15, 16]. For example, Ma *et al.* have introduced the first exploration of Ad Hoc multimedia event detection when there are only 10 positive examples for training [15]. However, the area of research remains in its infancy, thus motivating us to ask for more satisfying performance. As complex events usually contain visual attributes related to people, scenes, objects and human actions (*e.g.*, Figure 1 shows that a complex event *parkour* is relevant with *push ups*, *building*, *etc.*), leveraging these attributes properly could be helpful for the detection.

Visual attributes were introduced as describable properties of an object and have been applied to many applications [5, 7, 9]. Visual attributes can be either at local level

1063-6919/13 $26.00 © 2013 IEEE
DOI 10.1109/CVPR.2013.339

or global level. For example, *"many people"* is a local-level attribute for the event *flash mob gathering*. Yet this kind of attributes is mostly defined manually, which is time-consuming and requires expertise. Instead, we can use attributes at global level which are the semantic labels of images [19]. For instance, an image with its semantic label *tennis* can be leveraged for understanding the event *playing tennis*. Given that this type of attributes is associated with images, we regard it as image attributes. Using image attributes for complex event detection is intuitively limited as image attributes usually cannot characterize the dynamic properties of complex event videos (complex event videos refer to the videos depicting complex events). In this paper, we therefore propose an idea of video attributes and particularly apply it for complex event detection. **Video attributes**, in our work, indicate the semantic labels of other external videos collected by researchers. Note that these external videos are different from complex event videos. Compared to complex event videos, the external videos contain simple contents of people, objects, scenes and actions which are basic elements of complex events. For example, a video with its semantic label *mixing batter* is useful for understanding the complex event *making a cake*. As the external videos are used by treating their semantic labels as video attributes, we call these videos **attribute videos**.

To use video attributes, we may refer to a typical approach that involves training attribute classifiers and then using their outputs as intermediate representations for the complex event videos [8, 11]. But this approach has two problems. First, when the number of attributes used is limited, it is insufficient to learn a discriminative intermediate representation. Second, given a particular event to detect, only some attributes are discriminative while others are comparatively useless or even noisy [16]. It is difficult to decide what attributes to use for different events. In contrast, we propose to use video attributes as additional information to assist complex event detection. Specifically, our framework learns the attribute classifier and event detector simultaneously. The observation of a particular event affects the attribute classifier, and in return, attributes characterize the event. This kind of mutual influence is explored by a correlation vector, which helps incorporate extra informative cues into the event detector. We name the proposed method Multi-level Collaborative Regression (MCR). Our approach has two merits: the learning process of event detector is not solely dependent on the video attributes; and the joint framework adapts the knowledge from attributes for different events, *i.e.*, a particular event obtains dedicated perks via the joint learning of attribute classifier and event detector.

Moreover, we propose to integrate multiple features from both complex event videos and attribute videos for learning the detector as combining multiple features has proved to be beneficial for visual analysis [22]. On the other hand, existing video collections have different themes. As we expect the video attributes to be diverse, in our framework video attributes from different collections are utilized. To this end, we illustrate our approach in Figure 1.

The main contributions of this paper are as follows: First, we propose using video attributes for complex event detection. Second, video attributes are used latently as additional information for learning the event detector. Third, multiple attribute video sets with different features are sewed seamlessly with multiple features from the complex event videos.

2. Related Work

Visual attributes were advocated as the describable properties of objects [8]. For example, an object *bear* can be described by attributes such as *furry* and *four legs*. Attributes are both machine-detectable and human-understandable, so they have been widely used for various applications. Wang *et al.* have proposed a discriminative model for object recognition [21]. The attributes of an object are treated as latent variables and the correlations among attributes are used to classify object classes. A method to learn visual attributes and object classes together has been presented in [20]. Duan *et al.* have presented an interactive approach which discovers local attributes that are both discriminative and semantically meaningful for fine-grained category classification [7]. Hwang *et al.* have proposed to explore the shared features between objects and their attributes for animal and scene classification [9]. Dhar *et al.* have leveraged high level describable attributes for selecting high aesthetic quality images and interesting ones from large image collections [5]. However, to generate local attributes usually requires a manually defining process which is burdensome. An alternative way is to leverage attributes at a global level, *i.e.*, the semantic labels of visual data. Its convenience is that we have many labeled datasets covering a wide range of themes. By treating the semantic labels of these data as attributes, we can readily leverage them.

In the past, global-level image attributes have been widely used [19, 14]. For example, Luo *et al.* have presented an object classification method by casting prior features obtained from global image attributes of auxiliary images into their multiple kernel learning framework [14]. For recognition or detection tasks in videos, image attributes probably cannot well characterize the dynamic properties which could hamper their contributions.

In [16], Ma *et al.* have proposed learning an intermediate representation for event detection. In their approach, the intermediate representation is the same for the event videos and the attribute videos. However, it could be natural to assume that the events and attributes are different depictions of videos at different levels. In addition, the intermediate representation is unexplainable. Differently, we leverage

video attributes to characterize complex events, which is interpretable. In our framework, we also learn an attribute classifier which can be used to predict the attributes of a given video. Since the attribute classifier is jointly optimized with the event detector, the related attribute classifier is more accurate in uncovering the attributes from an event video. For example, by exploiting the videos of "landing a fish", the concept classifier "fish" can be more accurately trained and vice versa. In addition, as a byproduct of our method, the attribute representation can be further used for other applications such as multimedia event recounting [6].

3. Video Attributes Assisted Event Detection

We first correlate the features of attribute videos from m multiple sources with their semantic labels respectively. The features of different sources can be different. Following [15, 16], we perform full rank principal component analysis [18] to map the features into a Hilbert space $\mathcal{H}$. Denote their representations in $\mathcal{H}$ as $\tilde{V}_i\Big|_{i=1}^{m} \in \mathbb{R}^{d_i \times n_i}$ where d_i is the dimension and n_i indicates the number of videos. Suppose the semantic labels are $A_i|_{i=1}^{m} \in \mathbb{R}^{n_i \times c_i}$ where c_i is the number of classes, we propose the following regression loss:

$$\min_{Q_i} \sum_{i=1}^{m} \left\| \tilde{V}_i^T Q_i - A_i \right\|_F^2, \tag{1}$$

where $Q_i \in \mathbb{R}^{d_i \times c_i}$ associates $\tilde{V}_i$ with A_i. Next we illustrate how to learn a detector for the complex event by incorporating the attribute videos. Similarly we first map the multiple features of the complex event videos into $\mathcal{H}$ and denote the resulted representations as $\tilde{X}_i\Big|_{i=1}^{m} \in \mathbb{R}^{d_i \times n}$, where n is the number of complex event videos. We first propose to learn multiple detectors $w_i \in \mathbb{R}^{d_i \times 1}$ to associate $\tilde{X}_i$ with the ground truth labels $y \in \mathbb{R}^{n \times 1}$:

$$\min_{w_i} \sum_{i=1}^{m} \left\| \tilde{X}_i^T w_i - y \right\|_2^2. \tag{2}$$

On top of the above function, we aim to correlate different feature types in a joint framework. It is expected that the learning process from different feature types is sewed seamlessly to obtain better w_i. Hence, we bring in the predicted labels $f_i \in \mathbb{R}^{n \times 1}$ for each feature type and minimize the following objective:

$$\min_{w_i, f_i} \sum_{i=1}^{m} \left\| \tilde{X}_i^T w_i - f_i \right\|_2^2 + \| f_i - y \|_2^2. \tag{3}$$

Now we show how to incorporate the attribute videos for optimizing w_i. Since the attribute videos and the complex event videos are relevant, *i.e.*, complex events are usually related to people, scenes, objects and human actions, the two domains would have some shared knowledge. Inspired by previous works [4], we assume that a correlation vector $p_i \in \mathbb{R}^{c_i \times 1}$ exists to establish the correspondence between Q_i and w_i. Thus, Eq (3) is extended as:

$$\min_{w_i, Q_i, p_i, f_i} \sum_{i=1}^{m} \left\| \tilde{X}_i^T (w_i + \beta Q_i p_i) - f_i \right\|_2^2 + \| f_i - y \|_2^2, \tag{4}$$

where β is a parameter to control the influence of the attribute videos on the event detection. To this end, our objective function is formulated as follows:

$$\begin{aligned} &\min_{w_i, Q_i, p_i, f_i} \sum_{i}^{m} \left\| \tilde{V}_i^T Q_i - A_i \right\|_F^2 \\ &+\alpha \left(\left\| \tilde{X}_i^T (w_i + \beta Q_i p_i) - f_i \right\|_2^2 + \| f_i - y \|_2^2 \right) + \gamma \| w_i \|_2^2, \end{aligned} \tag{5}$$

where the last item is added to avoid over-fitting. For a testing video, $(w_i + \beta Q_i p_i)$ is used for prediction.

4. Optimization Procedure

We propose an alternating approach to optimize the objective function in Eq (5).

First, we fix p_i and optimize f_i, w_i and Q_i. By setting the derivative of Eq (5) w.r.t. f_i to zero, we have:

$$f_i = \left(\tilde{X}_i^T (w_i + \beta Q_i p_i) + y \right) / 2. \tag{6}$$

Substituting Eq (6) into Eq (5) we obtain:

$$\begin{aligned} &\min_{w_i, Q_i} \sum_{i}^{m} \left\| \tilde{V}_i^T Q_i - A_i \right\|_F^2 + \alpha \left\| \tilde{X}_i^T (w_i + \beta Q_i p_i) - y \right\|_2^2 \\ &+\gamma \| w_i \|_2^2. \end{aligned} \tag{7}$$

By setting the derivative of Eq (7) w.r.t. w_i to zero, it becomes:

$$w_i = \alpha B_i^{-1} \tilde{X}_i y - \alpha \beta B_i^{-1} \tilde{X}_i \tilde{X}_i^T Q_i p_i \tag{8}$$

where $B_i = \alpha \tilde{X}_i \tilde{X}_i^T + \gamma I$. Substituting Eq (8) into Eq (7) we have:

$$\begin{aligned} &\min_{Q_i} \sum_{i}^{m} \left\| \tilde{V}_i^T Q_i - A_i \right\|_F^2 + \alpha Tr(2\alpha\beta y^T \tilde{X}_i^T B_i^{-1} \tilde{X}_i \tilde{X}_i^T Q_i p_i \\ &-\alpha\beta^2 p_i^T Q_i^T \tilde{X}_i \tilde{X}_i^T B_i^{-1} \tilde{X}_i \tilde{X}_i^T Q_i p_i + \beta^2 p_i^T Q_i^T \tilde{X}_i \tilde{X}_i^T Q_i p_i \\ &-2\beta p_i^T Q_i^T \tilde{X}_i y). \end{aligned} \tag{9}$$

By setting the derivative of Eq (9) w.r.t. Q_i to zero, we arrive at:

$$\begin{aligned} &2\tilde{V}_i \tilde{V}_i^T Q_i - 2\tilde{V}_i A_i + 2\alpha^2 \beta \tilde{X}_i \tilde{X}_i^T B_i^{-1} \tilde{X}_i y p_i^T \\ &-2\alpha^2 \beta^2 \tilde{X}_i \tilde{X}_i^T B_i^{-1} \tilde{X}_i \tilde{X}_i^T Q_i p_i p_i^T + 2\beta^2 \tilde{X}_i \tilde{X}_i^T Q_i p_i p_i^T \\ &-2\beta \tilde{X}_i y p_i^T = 0 \end{aligned} \tag{10}$$

which can be rewritten as:

$$\begin{aligned} &Q_i(p_i p_i^T)^{-1} + \Big(\beta^2(\tilde{V}_i\tilde{V}_i^T)^{-1}\tilde{X}_i\tilde{X}_i^T - \alpha^2\beta^2(\tilde{V}_i\tilde{V}_i^T)^{-1} \\ &\tilde{X}_i\tilde{X}_i^T B_i^{-1}\tilde{X}_i\tilde{X}_i^T\Big) Q_i - (\tilde{V}_i\tilde{V}_i^T)^{-1}\tilde{V}_i A_i (p_i p_i^T)^{-1} \\ &+\alpha^2\beta(\tilde{V}_i\tilde{V}_i^T)^{-1}\tilde{X}_i\tilde{X}_i^T B_i^{-1}\tilde{X}_i y p_i^T (p_i p_i^T)^{-1} \\ &-\beta(\tilde{V}_i\tilde{V}_i^T)^{-1}\tilde{X}_i y p_i^T (p_i p_i^T)^{-1} = 0. \end{aligned} \tag{11}$$

The above problem can be solved by the Sylvester equation [1].

After Q_i, w_i and f_i are obtained, we fix them and optimize p_i. By setting the derivative of Eq (5) w.r.t. p_i to zero, we have:

$$p_i = (\beta Q_i^T \tilde{X}_i\tilde{X}_i^T Q_i)^{-1}(Q_i^T\tilde{X}_i f_i - Q_i^T\tilde{X}_i\tilde{X}_i^T w_i). \tag{12}$$

Thereby, we propose the algorithm shown in Algorithm 1 to optimize the objective function in Eq (5).

Algorithm 1: Optimization procedure for MCR.

Input:
$\tilde{V}_i \in \mathbb{R}^{d_i \times n_i}$, $A_i \in \mathbb{R}^{n_i \times c_i}$, $\tilde{X}_i \in \mathbb{R}^{d_i \times n}$, $y \in \mathbb{R}^{n\times 1}$; Parameters α, β and γ.

Output:
Optimized $w_i \in \mathbb{R}^{d_i\times 1}$, $Q_i \in \mathbb{R}^{d_i\times c_i}$, $p_i \in \mathbb{R}^{c_i\times 1}$ and $f_i \in \mathbb{R}^{n\times 1}$.

1: Set $t = 0$ and initialize $p_i \in \mathbb{R}^{c_i\times 1}$ randomly;
2: **repeat**
  Compute f_i according to Eq (6);
  Compute w_i according to Eq (8) ;
  Solve the Sylvester equation in Eq (11) to get Q_i;
  Update p_i according to Eq (12);
  $t = t + 1$.
  until *Convergence:* $|obj_{t+1} - obj_t|\,/obj_t \leq 10^{-3}$ *(obj indicates the objective function value)*;
3: Return w_i, Q_i, p_i and f_i.

5. Experiments

In this section we present the experiments that evaluate the proposed method for complex event detection.

5.1. Datasets

The TRECVID MED 2012 development set (MED12) is used for complex event detection. MED12 consists of 50328 video clips which are related to 20 events: *Birthday party*, *Changing a vehicle tire*, *Flash mob gathering*, *Getting a vehicle unstuck*, *Grooming an animal*, *Making a sandwich*, *Parade*, *Parkour*, *Repairing an appliance*, *Working on a sewing project*, *Attempting a bike trick*, *Cleaning an appliance*, *Dog show*, *Giving directions to a location*, *Marriage proposal*, *Renovating a home*, *Rock climbing*, *Town hall meeting*, *Winning a race without a vehicle* and *Working on a metal crafts project*.

Another two video sets, *i.e.*, the UCF50 dataset [17] and the development set from TRECVID 2012 semantic indexing task are used as attribute videos. UCF50 includes 6681 video sequences with 50 action categories. The video set for TRECVID 2012 semantic indexing (SIN) task covers 346 concepts. We use 65 concepts suggested by [6]. These concepts are related to human, scenes and objects which are the elements of events. The sampled subset contains 3244 data and we denote it as SIN12.

We extract STIP [12] and SIFT [13] descriptors for the videos of MED12, STIP for UCF50 and SIFT for SIN12. After that, a 32768 dimension spatial BoW feature is formed for STIP/SIFT to represent each video.

5.2. Comparison Algorithms

(1) MCR: The proposed method in this paper. As χ^2 kernel has proved to be advantageous for BoW feature, we exploit it to map the features of MED12, UCF50 and SIN12 into the Hilbert space.

(2) Baseline: We set β in Eq (5) to 0 so that no video attributes are exploited in our approach. The resulting algorithm works as the baseline.

(3) SVM: SVM is an effective tool for complex event detection and has been widely used by several research groups for TRECVID MED, *e.g.*, [23]. Similarly, χ^2 kernel is used.

(4) Attributes Intermediate Representation (AIR): We train attribute classifiers using UCF50 and SIN12. Then we apply the classifiers on MED12 and use their outputs as the intermediate representations. SVM is applied on the new representations afterwards for event detection.

5.3. Setup

For each event, we randomly choose 100 positive examples and 1000 negative examples from MED12 to form the training set. The remaining data of MED12 are used as the testing set.

There are two types of parameters. The first type includes the parameters for kernel calculation. It is fixed to the mean of the pairwise distances among the training samples as done in [14]. The second type includes the regularization parameters. We tune them uniformly from $\{0.001, 0.1, 10, 1000\}$ for all the algorithms and we report the best results for each algorithm.

We use three evaluation metrics. Minimum NDC (MinNDC) and the Probability of Miss-Detection based on the Detection Threshold 12.5 (Pmd@12.5) are two official evaluation metrics used by NIST in TRECVID MED [2][3]. Lower MinNDC or Pmd@12.5 indicates better detection performance. The third one is Average Precision (AP). Higher AP indicates better performance.

Table 1. Detection results using different algorithms. LOWER MinNDC / LOWER Pmd@12.5 / HIGHER AP indicates BETTER performance. The best results are highlighted in bold. Relative Improvement indicates our advantage over the runner-up, if applicable.

Event Description	Evaluation Metric	Baseline	SVM	AIR	MCR	Relative Improvement
Birthday party	MinNDC	0.900	0.877	1.000	**0.858**	2.2%
	Pmd@12.5	0.516	0.498	0.989	**0.484**	2.9%
	AP	0.064	0.068	0.007	**0.076**	11.7%
Changing a vehicle tire	MinNDC	0.895	0.753	1.000	**0.719**	4.7%
	Pmd@12.5	0.529	0.443	0.979	**0.436**	1.6%
	AP	0.032	0.058	0.003	**0.069**	19.0%
Flash mob gathering	MinNDC	0.463	0.467	0.721	**0.420**	10.2%
	Pmd@12.5	0.239	0.249	0.394	**0.230**	3.9%
	AP	0.225	0.225	0.087	**0.248**	10.2%
Getting a vehicle unstuck	MinNDC	0.710	0.607	0.957	**0.559**	8.6%
	Pmd@12.5	0.391	**0.326**	0.710	0.355	N/A%
	AP	0.071	0.095	0.014	**0.118**	24.2%
Grooming an animal	MinNDC	0.935	0.908	1.000	**0.855**	6.2%
	Pmd@12.5	0.532	**0.511**	0.957	**0.511**	N/A
	AP	0.026	0.029	0.003	**0.034**	17.2%
Making a sandwich	MinNDC	0.950	0.905	0.985	**0.888**	1.9%
	Pmd@12.5	0.546	0.540	0.741	**0.517**	4.4%
	AP	0.032	0.037	0.012	**0.039**	5.4%
Parade	MinNDC	0.761	0.747	0.991	**0.683**	9.4%
	Pmd@12.5	0.391	0.407	0.579	**0.374**	4.5%
	AP	0.123	0.124	0.044	**0.141**	13.7%
Parkour	MinNDC	0.610	0.576	0.878	**0.534**	7.9%
	Pmd@12.5	0.384	**0.344**	0.528	**0.344**	N/A
	AP	0.092	0.108	0.030	**0.117**	8.3%
Repairing an appliance	MinNDC	0.728	0.689	0.935	**0.630**	9.4%
	Pmd@12.5	0.402	0.386	0.614	**0.378**	2.1%
	AP	0.064	0.066	0.019	**0.084**	27.3%
Working on a sewing project	MinNDC	0.817	0.753	0.964	**0.721**	4.4%
	Pmd@12.5	0.475	0.475	0.639	**0.459**	3.5%
	AP	0.042	0.042	0.015	**0.048**	14.3%
Attempting a bike trick	MinNDC	0.692	**0.556**	1.000	0.559	N/A
	Pmd@12.5	0.433	**0.333**	0.800	**0.333**	N/A
	AP	0.015	0.022	0.001	**0.025**	13.6%
Cleaning an appliance	MinNDC	0.978	0.957	1.000	**0.852**	12.3%
	Pmd@12.5	0.600	0.700	0.900	**0.467**	28.5%
	AP	0.005	0.004	0.001	**0.007**	40.0%
Dog show	MinNDC	0.545	0.434	0.943	**0.390**	11.3%
	Pmd@12.5	0.300	0.267	0.600	**0.200**	33.5%
	AP	0.028	0.037	0.004	**0.043**	16.2%
Giving directions to a location	MinNDC	0.862	0.875	1.000	**0.844**	3.7%
	Pmd@12.5	0.670	**0.667**	0.967	**0.667**	N/A
	AP	0.005	0.005	0.001	**0.006**	20%
Marriage proposal	MinNDC	0.824	**0.774**	1.000	0.777	N/A
	Pmd@12.5	0.533	**0.500**	0.967	**0.500**	N/A
	AP	0.008	**0.011**	0.001	**0.011**	N/A
Renovating a home	MinNDC	0.821	0.821	1.000	**0.735**	11.7%
	Pmd@12.5	0.567	0.533	0.867	**0.467**	14.1%
	AP	0.008	0.009	0.001	**0.013**	44.4%
Rock climbing	MinNDC	0.659	0.670	0.949	**0.575**	14.6%
	Pmd@12.5	0.431	0.433	0.633	**0.400**	7.8%
	AP	0.017	0.016	0.004	**0.023**	35.3%
Town hall meeting	MinNDC	0.706	0.607	1.000	**0.532**	14.1%
	Pmd@12.5	0.467	0.367	1.000	**0.300**	22.3%
	AP	0.016	**0.023**	0.007	0.020	N/A
Winning a race without a vehicle	MinNDC	0.683	0.585	0.887	**0.565**	3.5%
	Pmd@12.5	0.433	**0.333**	0.667	0.367	N/A
	AP	0.018	0.021	0.004	**0.023**	9.5%
Working on a metal crafts project	MinNDC	0.822	0.750	0.947	**0.690**	8.7%
	Pmd@12.5	0.500	**0.400**	0.633	**0.400**	N/A
	AP	0.009	0.012	0.004	**0.018**	50.0%
Average	MinNDC	0.768	0.716	0.958	**0.669**	7.0%
	Pmd@12.5	0.467	0.436	0.758	**0.409**	6.6%
	AP	0.045	0.053	0.013	**0.061**	15.1%

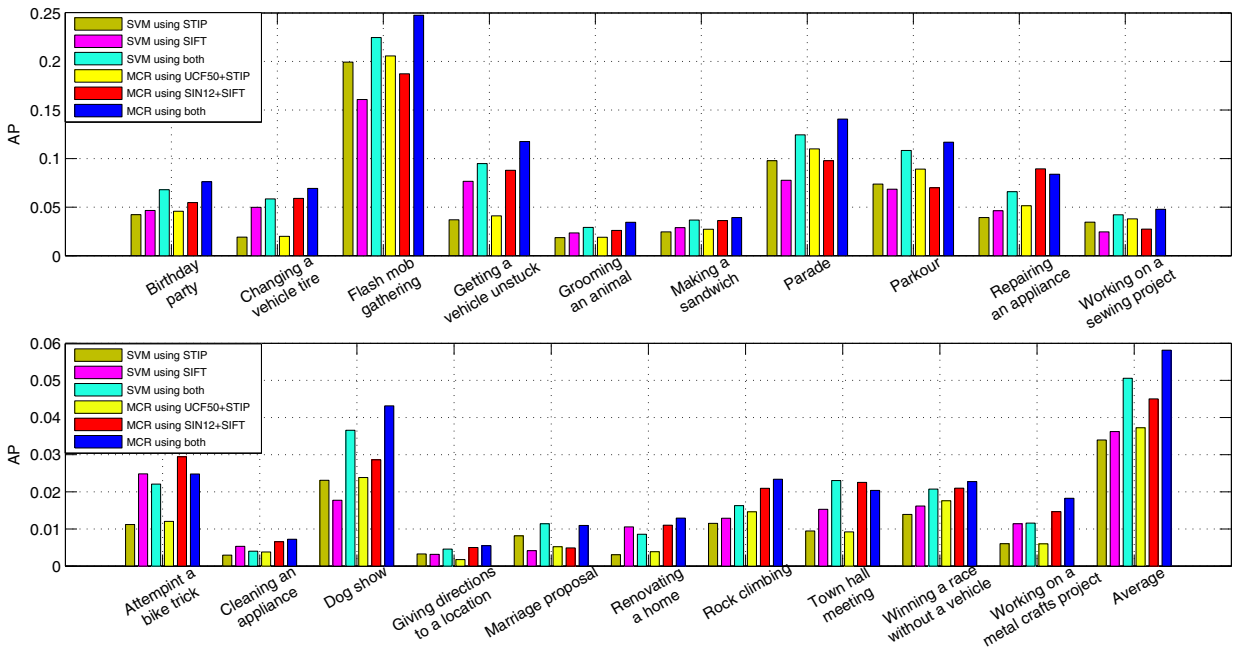

Figure 2. Performance variation w.r.t. feature type and source attribute.

5.4. Event Detection Results

Table 1 lists the detection results. It can be seen that our method MCR is consistently competitive for all the events. Specifically, we observe that: 1) when using MinNDC and Pmd@12.5 as metrics, MCR gains the best performance for 18 events; 2) when using AP as metric, MCR is the best method for 19 events; 3) MCR obtains the top performance for the average accuracy over all the 20 events; 4) MCR is much better than the Baseline, indicating that harnessing video attributes does boost the performance of complex event detection; 5) SVM is the second competitive algorithm, which is in accordance with previous experience of several research groups in TRECVID MED; 6) for those events on which MCR achieves the top performance, it outperforms SVM with clear gap. For instance, MCR is 10%-75% better than SVM for 16 events in terms of AP. The promising performance of MCR verifies that leveraging video attributes properly is beneficial for complex event detection.

5.5. Results using Single Feature and Single Source

In this part, we only use UCF50+MED12 with STIP feature and SIN12+MED12 with SIFT feature for complex event detection to show the performance change. As SVM is the second competitive algorithm, we also show its performance variation w.r.t. STIP feature and SIFT feature. Due to the space limit, we only show the results using AP as metric for this experiment. The results are displayed in Figure 2. It is observed that: 1) MCR using both UCF50 and SIN12 together with STIP+SIFT features is better than that using UCF50+MED12 with STIP feature for all the events; 2) MCR using both UCF50 and SIN12 together with STIP+SIFT features is generally better than that using SIN12+MED12 with SIFT feature, yet the former is weaker than the latter for three events, which is presumably data-dependent; 3) SVM has similar performance variation; and 4) our method still yields better results than SVM when using one feature type. This experiment validates that exploiting multiple attribute video sets together with different features is beneficial for most cases.

6. Conclusions

We have proposed a method for utilizing the attributes at video level for complex event detection. Video attributes are convenient to use for complex event detection as many video collections relevant to people, scenes, objects and actions are available. Meanwhile, video attributes have more potentials than image attributes to characterize the dynamic properties of video data. Unlike the traditional approach which maps the video data into attribute space, our method learns a correlation vector which correlates video attributes and a complex event. Built upon this, the extra informative cues learnt from attribute videos are further incorporated into the event detector. We have performed extensive experiments using a real-world large-scale video dataset to evaluate the efficacy of our method on complex event detection. The results are encouraging and have verified the advantage of leveraging video attributes properly.

7. Acknowledgments

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

This paper was partially supported by the S-PATTERNS FIRB project, the Singapore National Research Foundation under its International Research Centre @ Singapore Funding Initiative and administered by the IDM Programme Office, and the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center contract number D11PC20068. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, or the U.S. Government.

References

[1] http://en.wikipedia.org/wiki/sylvester_equation.

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

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

[4] R. K. Ando and T. Zhang. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6:1817–1853, 2005.

[5] S. Dhar, V. Ordonez, and T. L. Berg. High level describable attributes for predicting aesthetics and interestingness. In *CVPR*, pages 1657–1664, 2011.

[6] D. Ding, F. Metze, S. Rawat, P. F. Schulam, S. Burger, E. Younessian, L. Bao, M. G. Christel, and A. G. Hauptmann. Beyond audio and video retrieval: towards multimedia summarization. In *ICMR*, 2012.

[7] K. Duan, D. Parikh, D. J. Crandall, and K. Grauman. Discovering localized attributes for fine-grained recognition. In *CVPR*, pages 3474–3481, 2012.

[8] A. Farhadi, I. Endres, D. Hoiem, and D. A. Forsyth. Describing objects by their attributes. In *CVPR*, pages 1778–1785, 2009.

[9] S. J. Hwang, F. Sha, and K. Grauman. Sharing features between objects and their attributes. In *CVPR*, pages 1761–1768, 2011.

[10] H. Izadinia and M. Shah. Recognizing complex events using large margin joint low-level event model. In *ECCV (4)*, pages 430–444, 2012.

[11] 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.

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

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

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

[15] 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*, pages 469–478, 2012.

[16] Z. Ma, Y. Yang, N. Sebe, K. Zheng, and A. G. Hauptmann. Multimedia event detection using a classifier-specific intermediate representation. *IEEE Transactions on Multimedia*, 2013.

[17] K. K. Reddy and M. Shah. Recognizing 50 human action categories of web videos. *Machine Vision and Applications Journal*, 2012.

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

[19] L. Torresani, M. Szummer, and A. W. Fitzgibbon. Efficient object category recognition using classemes. In *ECCV (1)*, pages 776–789, 2010.

[20] G. Wang and D. A. Forsyth. Joint learning of visual attributes, object classes and visual saliency. In *ICCV*, pages 537–544, 2009.

[21] Y. Wang and G. Mori. A discriminative latent model of object classes and attributes. In *ECCV (5)*, pages 155–168, 2010.

[22] Y. Yang, J. Song, Z. Huang, Z. Ma, N. Sebe, and A. G. Hauptmann. Multi-feature fusion via hierarchical regression for multimedia analysis. *IEEE Transactions on Multimedia*, 2013.

[23] S.-I. Yu, Z. Xu, D. Ding, W. Sze, F. Vicente, Z. Lan, Y. Cai, S. Rawat, P. Schulam, N. Markandaiah, S. Bahmani, A. Juarez, W. Tong, Y. Yang, S. Burger, F. Metze, R. Singh, B. Raj, R. Stern, T. Mitamura, E. Nyberg, and A. Hauptmann. Informedia e-lamp @ TRECVID2012: Multimedia event detection and recounting med and mer. In *NIST TRECVID Workshop*, 2012.