1-2016

Object pooling for multimedia event detection and evidence localization

Ho ZHANG

Chong-Wah NGO

Chong-wah NGO
Singapore Management University, cwngo@smu.edu.sg

Invited Paper

Object Pooling for Multimedia Event Detection and Evidence Localization

Hao Zhang †, Chong-Wah Ngo †

Abstract Multimedia event detection (MED) and evidence hunting are two primary topics in the area of multimedia event search. The former serves to retrieve a list of relevant videos given an event query, whereas, the latter reasons why and how much the degree a retrieved video answers that query. Common practices deal with these two topics in separate methods, however, in this paper, we combine MED and evidence hunting into a joint framework. We propose a refined semantical representation named object pooling which can dynamically extract visual snippets corresponding to the location of when and where evidences might appear. The main idea of object pooling is to adaptively sample regions from frames for generation of object histogram that can be efficiently rolled up and back. Experiments conducted on large-scale TRECVID MED 2014 dataset demonstrate the effectiveness of proposed object pooling approach on both event detection and evidence hunting.

Key words: Object Pooling, Event Modeling, Search Result Reasoning

1. Introduction

With the popularity of video sharing website, such as YouTube, thousands of user-generated videos are uploaded onto Internet every day. Unlike professionally filmed videos (e.g., sport videos), these videos usually suffer from low resolution, large diversities and reflect complex event contents. As showed in Figure 1, the sport "swimming" contains single interaction "people swim in water" captured from different camera views, whereas, multimedia event "birthday party" usually contains a sequence of interactions, such as "people sing birthday song", "people blow candles", "people eat cake" and "people receive gift" in different scenarios. Additionally, single sport video usually contains significant motion patterns with few irrelevant clips, on the contrary, multimedia video contains complex motions with many irrelevant clips. As a result, the method of sport detection with visual features (e.g., HOG, SIFT, Improved Dense Trajectory) is not quite appropriate for multimedia event detection and summarization, bringing the needs to capture semantical meanings in multimedia videos. To efficiently retrieve and summarize semantical meanings of multimedia videos, many research efforts have been spent on representing multimedia contents with high-level semantical concepts. For example,

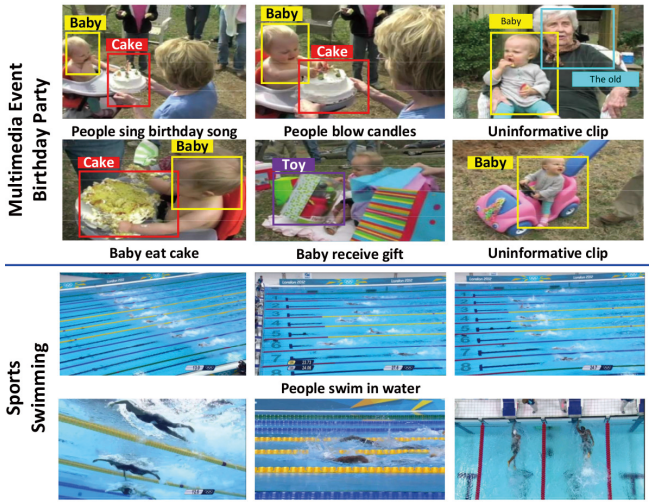

Fig. 1 Comparison of Multimedia Event and Sport Event.

the event "birthday party" usually contains relevant objects such as "people", "birthday cake", "candles", "toy gift", etc (Figure 1). The video snippets having these objects and/or actions could be extracted as evidences for justifying the presence of a target event. As a result, multimedia event retrieval can significantly benefit from the effective semantic representation of web videos. However, the state-of-the-art concept detectors are trained with deep convolutional neural networks which yield high responses for primary objects occupying large frame-area and downgrade responses for regional objects covering small area, as a result, regional objects are usually neglected.

The main contribution of this paper is on the proposal of a refined semantical concept representation

Received November 29, 2015; Accepted January 6, 2016

† City University of Hong Kong.
(Hong Kong, China)

for multimedia videos by using an *object pooling* approach. The object evidences are locally extracted out of video frames, which are further pooled within and across frames to form an object histogram. The histogram can be directly utilized for retrieving event relevant videos. During event evidence localization, the histogram can be unrolled in time such that temporal evidences, or specifically visual snippets, in the videos can be readily identified in a unified way. Furthermore, the spatial regions of event related objects can also be effortlessly located through the "unrolled histograms" (see Figure 2). Our model is the first attempt to locate not only *when* (i.e., which video frames) evidential object appears, but also *where* (i.e., which spatial region) it resides. Additionally, with the help of homogeneous kernel mapping, we also propose a novel method to identify important elements of nonlinear kernel SVM (e.g., χ^2-SVM). The experimental results demonstrate that the proposed object pooling can both improve the performance of multimedia event detection and provide reasonable event evidences.

This paper investigates multimedia event detection and spatial-temporal localization of visual evidences in explaining video relevancy in a unified way by using a large pool of concept detectors. We restrict the studies to the domain of multimedia event, where each event usually has a number of objects interacting with each other.

The rest of this paper is organized as follows: Section 2 describes the Related Works on multimedia event detection and event evidences localization. Section 3 describes Object Pooling for multimedia event detection, while Section 4 describes Evidential Objects Localization, Section 5 presents Experimental Settings, and Sections 6 discusses Results. Section 7 concludes the paper.

2. Related Works

Detection and summarization are two primary problems in the area of multimedia video analysis. Content-based video retrieval detects event relevant videos, and, video summarization helps to verify why a retrieved video is being relevant. From the user point of view, providing such a feature, which is equivalent to on-the-fly generation of short summaries as evidences explaining how a video answers an event query (a.k.a. multimedia recounting[1]), can have potential in enhancing user search experience. More specifically, instead of watching throughout a video, user can rapidly determine the relevancy by simply reading evidences[2].

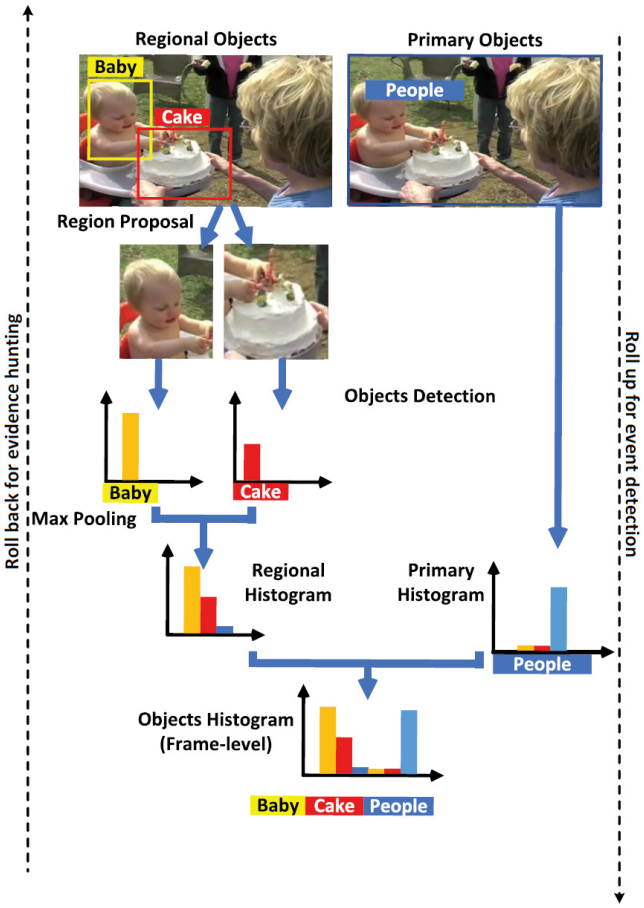

Fig. 2 Overview of Object Pooling across spatial domain. Primary objects are detected from frame and regional objects are accumulated from regions. The two kinds of object histograms are concatenated to represent a video frame.

In the literature, feature pooling in the spatial[3] and temporal[4][5][6] dimensions is a technique intensively studied. In the context of multimedia event, most pooling techniques involve accumulation of features along the temporal dimension. For example, the recent studies[7][8] employ multiple-instance like algorithms in estimating the importance of video frames in the pooling features. Our proposed work is different in the way that we mainly deal with object features at the frame level (see Figure 2). Compared with approaches such as spatial pyramid[3], which rigidly divides a frame into some predefined regions for spatial pooling, object pooling has the capability of adaptively sampling regions where objects are likely to reside. There are also few works, such as[9][10][11], focusing on the selection of few relevant concepts from large-scale concept sets for multimedia event detection and achieving promising results. Object pooling refines semantical representation and can be treated as the preprocessing of concept for selection.

There are some related works studying multimedia event recounting[2][12][13][14][15], but mostly treating video retrieval and recounting as two disintegrated parts. For

example in[13)14)15)], two different sets of classifiers, one for scoring video relevancy and the other for searching local evidences in a video, are separately trained. This strategy can lead to redundancy in training and detection, and more importantly, resulting in inconsistent judgements when scoring videos and hunting evidences. In contrast, object pooling provides a more sensible way of unifying video retrieval and recounting, by rolling up from local regions to a histogram for efficient event detection and drilling down the histogram for not only temporal but also spatial evidence localization, which are not yet studied in other approaches. Our approach has the advantage of simplicity and is easy to implement, if compared to other computationally expensive approaches such as[16)17)], where the former adopts DPM (deformable part based-model) requiring many training examples and the latter applies object classifiers (SVM) across an image at multi-scales which requires much computational cost.

Our work is motivated by the studies of object detection. With the help of deep convolutional neural networks (DCNN)[18)], the accuracy of object detection improves significantly. Girshick et al.[19)] proposed an object detection method called Region-CNN which utilizes DCNN features extracted from image regions to train object detector. Sermant et al.[20)] proposed OverFeat which applies DCNN classifier on image with multi-scale sliding windows. However, unlike these approaches which aim to assign region to object, this paper accumulates object responses from regions to generate a thorough representation for video frames.

3. Object Pooling

Multimedia event is usually generic in terms of event definition and complex in forms of audio-visual content. A feasible way of detecting events is by modeling the elementary concepts underlying an event, such as by representing an event as a histogram of concepts, which sometimes exhibits better performance than using low-level audio-visual features[16)21)22)23)24)]. Similar in spirit, we propose to represent each video frame by a refined semantical histogram which accumulates object responses from frames and frame regions.

Due to the existence of *soft-max* layer of deep convolutional neural networks, the state-of-the-art concept detectors (DCNN) are preferential to enhance response of primary object which occupies major region of an image, and penalize responses of regional objects with small areas which are informative for identifying the underlining events in the video. *Object pooling* is designed to address this issue by accumulating objects' responses from both frame and frame regions. This procedure starts by sampling the regions in video frames where objects may reside. A histogram of objects is then generated for each region, indicating the probability distribution of object appearance in that region. Then, the histograms of several regions are spatially pooled across regions to generate a frame-level representation (See Figure 2). Finally, we concatenate global and region representations for each frame. The resulting histogram is further temporally pooled along the temporal dimension. To this end, each video is represented by a video-level histogram serving as an input for SVM classifier learning.

3.1 Primary Objects

Primary objects usually locate at the center of an image occupying large areas, making them preferential to be detected and emphasized. By detecting concepts on each video frame, state-of-the-art methods mainly use primary objects to characterize contents of a video frame.

(1) Primary Object Detector

We employ *deep convolutional neural networks* (DCNN) for object detection. DCNN has been demonstrated to be effective in learning different levels of image representation and concept classifiers simultaneously.

Following standard pipeline, we apply DCNN classifiers to each frame and represent it with a vector of concept responses, where each element corresponds to the probability of appearance of an object. The primary object vector is denoted as $\boldsymbol{P}_t \in \mathbb{R}^D$, where D is the number of concepts and t denote the t-th frame.

3.2 Regional Objects

Regional objects are objects which reside in non-center parts of an image occupying small areas. When an image is input into concept detectors, responses of regional objects are usually overwhelmed by primary objects. However, regardless of their sizes, regional objects have potential to be event evidences, e.g., the "birthday cake" in the first video frame (Figure 1). Absences of regional objects will lead to incomplete representation of video contents. Thus, we propose to accumulate the semantical information of small regions by detecting the objects separately from primary objects.

(1) Region Proposal

Objects can appear at any locations of a frame. The

purpose of region proposal is to sample candidate windows, each of which contains an object. There are various off-the-shelf algorithms for determining the category independent object locations, e.g, selective search [25], objectness[26] and BING[27]. These algorithms are designed for images. For video frames that generally suffer from motion blur, we find that the color and brightness based segmentation algorithms such as selective search[25] are more appropriate than saliency and edge-based algorithms[26][27].

Given a video frame, we employ selective search to suggest candidate windows or bounding boxes for objects. The number of bounding boxes can be as many as a thousand. To save computational cost, we only process a small number of boxes which satisfy the predefined selection criteria. Basically, we minimize the overlapping area among the selected boxes, while pruning boxes of tiny size or with large ratio of width and height. For frame f_t extracted at timestamp t, we obtain K number of candidate regions denoted as, $B_t = \{b_{tk}\}_{k=1}^{K}$.

(2) Regional Object Detector

DCNN architecture can be used to detect and locate objects within a *part* of image, as demonstrated by object detection models such as OverFeat[20] and Region-CNN[19]. We use similar DCNN architecture proposed in[18] but fine-tuned with a large concept bank.

Given a video frame f_t and the proposed regions $B_t = \{b_{tk}\}_{k=1}^{K}$, we extract DCNN concept feature from each region. The feature corresponds to the neuronal response of DCNN given the visual content of a bounding box as input. Thus, each region b_{tk} is characterized as an object histogram $\boldsymbol{r}_{tk} \in \mathbb{R}^D$.

Spatial pooling is carried out by combining the region-level object histograms using the max^* operator. A frame-level histogram $\boldsymbol{R}_t \in \mathbb{R}^D$ is then generated as following:

$$\boldsymbol{R}_t = \ max([\boldsymbol{r}_{t1}, \quad \boldsymbol{r}_{t2}, \quad \cdots \quad , \boldsymbol{r}_{tK}]) \tag{1}$$

where $\boldsymbol{r}_{tk}$ denotes the object histogram for the k-th bounding box of the t-th frame.

3.3 Frame and Video Representation

(1) Frame-level Histogram

Each video frame is represented in two forms: primary object histogram and regional object histogram, where the former emphasizes main object with large size and the latter collects regional objects. We generate the frame-level histogram $\boldsymbol{h}_t \in \mathbb{R}^{2D}$ by concatenating the primary and regional representations:

$$\boldsymbol{h}_t = \Big[\boldsymbol{P}_t, \quad \boldsymbol{R}_t\Big] \tag{2}$$

(2) Video-level Histogram

The video-level histogram is generated by temporally pooling or rolling up histograms at the frame level using the *max* operator. Given a video of n subsampled frames, the histogram $\boldsymbol{H} \in \mathbb{R}^{2D}$ is obtained by using *max* (or *mean*[**]) operation:

$$\boldsymbol{H} = \ max([\boldsymbol{h}_1, \quad \boldsymbol{h}_2, \quad \cdots, \quad \boldsymbol{h}_n]) \tag{3}$$

4. Evidential Objects Localization

4.1 Event Detection

Given a training video set $\{\boldsymbol{H}_i, y_i\}_{i=1}^{V}$ with V videos, in which $\boldsymbol{H}_i$ is the feature, and y_i is the event label for the i-th video. An event detector is learned over training videos with nonlinear kernel SVM. The decision function of nonlinear kernel SVM for a testing video $\boldsymbol{H}$ is:

$$f(\boldsymbol{H}) = \boldsymbol{W}^T \boldsymbol{\Psi}(\boldsymbol{H}) \tag{4}$$

where $\boldsymbol{W}$ is the hyperplane learned by SVM and $\boldsymbol{\Psi}(\boldsymbol{H})$ is a nonlinear mapping function that projects vector $\boldsymbol{H}$ into a high dimensional space.

4.2 Evidence Localization

Before temporally and spatially locating key evidential objects, we need to identify which objects are evidences for events (e.g., "cake", "toy gift" for event "birthday"). We solve this problem by calculating and sorting contribution of each object for decision function $f(\boldsymbol{H})$. Objects which contribute most to the decision function $f(\boldsymbol{H})$ are identified as evidences.

An ideal situation to calculate contribution of each object is that the mapping function $\boldsymbol{\Psi}(\boldsymbol{H})$ can be applied to each element $\boldsymbol{H}(j)$ independently, thus, the contribution $C(j)$ of the j-th element $\boldsymbol{H}(j)$ can be calculated by below function:

$$C(j) = \boldsymbol{W}_j^T \boldsymbol{\Psi}\left(\boldsymbol{H}(j)\right) \tag{5}$$

where $\boldsymbol{W}_j^T$ is a subset of $\boldsymbol{W}^T$ corresponding to $\boldsymbol{\Psi}\left(\boldsymbol{H}(j)\right)$ in dimensions (see Figure 3). Thus, function 4 can be rewritten as:

$$f(\boldsymbol{H}) = \boldsymbol{W}^T \boldsymbol{\Psi}(\boldsymbol{H}) = \sum_{j=1}^{2D} C(j) \tag{6}$$

However, for nonlinear kernel, mapping function $\Psi(\boldsymbol{H})$

[*] If $\mathbf{A}$ is a matrix, $max(\mathbf{A})$ returns a column vector containing the maximum element from each row

[**] If $\mathbf{A}$ is a matrix, $mean(\mathbf{A})$ returns a column vector containing the mean value for each row

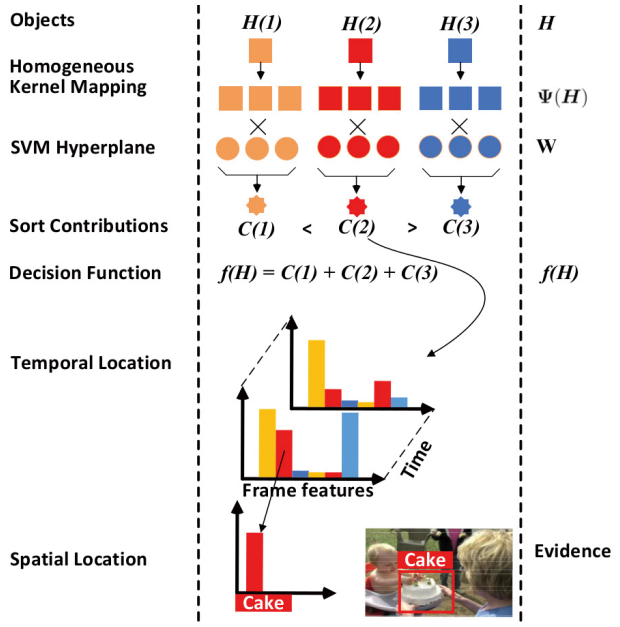

Fig. 3 Evidence Localization. Each dimension of object histogram is separately mapped into $\hat{\Psi}(x)$ space, then, the mapped feature is multiplied with corresponding SVM weights to calculate contribution. Object with large contribution $C(2)$ is identified as key evidence. By a reverse process of temporal-spatial pooling, *when* and *where* evidential object locate is identified.

usually has infinite dimension and elements of the mapped vector are not independent. The problem is solved by using a finite approximation of $\Psi(\boldsymbol{H})$ called homogeneous kernel mapping. In[28], Vedaldi et al. proposed homogeneous kernel mapping $\hat{\Psi}(\boldsymbol{H})$ to estimate $\Psi(\boldsymbol{H})$. Different from $\Psi(\boldsymbol{H})$, $\hat{\Psi}(\boldsymbol{H})$ can map each element $\boldsymbol{H}(j)$ independently into a $2m+1$ dimensional space. Thus, $\boldsymbol{H} \in \mathbb{R}^{2D}$ can be mapped into $2D(2m+1)$ dimensional space, where D denotes the number of concepts and m is a parameter for approximation function. In all, by utilizing homogeneous kernel mapping, we approximately calculate nonlinear kernel SVM using linear SVM in $\hat{\Psi}(\boldsymbol{H})$ space. There are various off-the-shelf nonlinear kernels, such as Hellinger, intersection and χ^2 kernels, which can be approximated by homogeneous kernel mapping. In this paper, we select χ^2-SVM since it demonstrates good performance. Considering computation cost, we only use $m = 1$ and map original feature into a 3 times dimensional space. The mapping function for each dimension is shown below.

$$\hat{\Psi}(\boldsymbol{H}(j)) = \sqrt{\boldsymbol{H}(j)} \begin{bmatrix} 0.8 \\ 0.6\cos(0.6\log \boldsymbol{H}(j)) \\ 0.6\sin(0.6\log \boldsymbol{H}(j)) \end{bmatrix} \quad (7)$$

With formulas (5, 7), $C(j)$ is calculated for the j-th object (i.e., the j-th element in $\boldsymbol{H}$). Then, we use $C(j)$ to rank contributions of objects for an multimedia event. Objects with top ranked contributions are identified as evidences.

ID	Event Name
E021	Attempting a bike trick
E022	Cleaning an appliance
E023	Dog show
E024	Giving directions to a location
E025	Marriage proposal
E026	Renovating a home
E027	Rock climbing
E028	Town hall meeting
E029	Winning a race without a vehicle
E030	Working on a metal crafts project
E031	Beekeeping
E032	Wedding shower
E033	Non-motorized vehicle repair
E034	Fixing a musical instrument
E035	Horse riding competition
E036	Felling a tree
E037	Parking a vehicle
E038	Playing fetch
E039	Tailgating
E040	Tuning a musical instrument

Table 1 The 20 events defined in TRECVID MED 2014 dataset

As is illustrated in Figure 3, we reverse the process of temporal-spatial pooling and relocate the important object in a spatial region of video frame, i.e., we locate *when* evidence appears by searching for the frame which has maximum object response and *where* evidence appears by searching for the region which has maximum object response.

5. Experimental Settings

5.1 Dataset

We conduct experiments using TRECVID MED 2014 dataset[1], which defines 20 multimedia events on three subsets: training, background and testing set. Different from pre-segmented action recognition dataset such as UCF-101, these videos are user-generated and not segmented into shots. The average video length is 2.4 minutes. The events include "attempting a bike trick", "dog show" and etc. Table 1 shows the complete list of events. In the training set, there are 1,996 positive videos in total, approximately 100 positives for each event. The background set has 4,992 randomly sampled videos may or may not be relevant to the twenty events. We treat this set as negative set. The testing set contains 27,276 videos independent from the other sets.

5.2 Concepts and Latent Concepts

We train concept detectors with DCNN architecture [18)34)]. Specifically, the DCNN architecture is implemented by Caffe[29] and pre-trained with ILSVRC-2012 data set[30] containing 1.26 million training images of 1000 categories.

To fully capture the semantical meaning of multimedia event, we utilize TRECVID SIN, MED Research Collection[1], ImageNet[30] dataset to create a large-scale *concept-bank* with 1,843 concepts. We summarize the

pipeline to generate concept detectors as below:

SIN-346: 346 concept detectors are fine-tuned on TRECVID SIN'14 dataset[1] with AlexNet structure[18]. Objects responses extracted by this DCNN are named as SIN-346.

RC-497: Similar to[31], we select 497 concepts from TRECVID MED'14 Research Collection[1], manually annotate at most 200 positive keyframes for each concept, and fine-tune 497 concepts using AlexNet structure. Similarly, this feature is named as RC-497.

ImageNet-1000: 1000 concept detectors are trained with AlexNet structure on a subset of ImageNet dataset [30] containing 1.26 million training images. For simplicity, the feature is named as ImageNet-1000.

Latent-Concept Descriptors: Compared to output layer (soft-max layer), even though the elements of fc7_relu feature (i.e., 4,096 dimensional outputs of the 2nd fully connected layer) are not assigned to human-understandable semantical labels, they can be used as visual descriptors reflecting semantical attributes. To differentiate fc7_relu feature with conceptual features, we refer the fc7_relu feature as latent-concept descriptors. We directly use VGG-19-layers Net[34] provided by Caffe and pre-trained with ImageNet dataset to extract latent-concept descriptors.

5.3 Regions

We use *"single strategy"* of selective search to propose around 200 alternative regions for each video frame. Specifically, single strategy segments objects based on color, size and texture similarity in HSV space.

To reduce computation cost, we remove tiny regions with areas smaller than $\frac{1}{10}$ of the area of whole frame. We also empirically remove the regions with $\frac{hight}{width}$ or $\frac{width}{hight}$ ratio larger than 4. Additionally, we prune bounding boxes with large overlaps. The overlap ratio between two regions (b_1 and b_2) is defined as: $overlap = \frac{area(b_1 \cap b_2)}{area(b_1 \cup b_2)}$. If two regions are overlapped by 0.5, we only retain one of them. By pruning these regions, we approximately obtain 20 regions for each video frame.

Since DCNN requires a fixed input size, we resize each frame region into the required resolution (227×227 for AlexNet, 224×224 for VGG-19-layers Net) first. Then, for each frame region, its representation is obtained by feeding it into pre-trained DCNN architecture. Thus, each region is characterized by DCNN outputs reflecting concept probabilities. Region-level features are further *max* pooled to generate the frame-level representation.

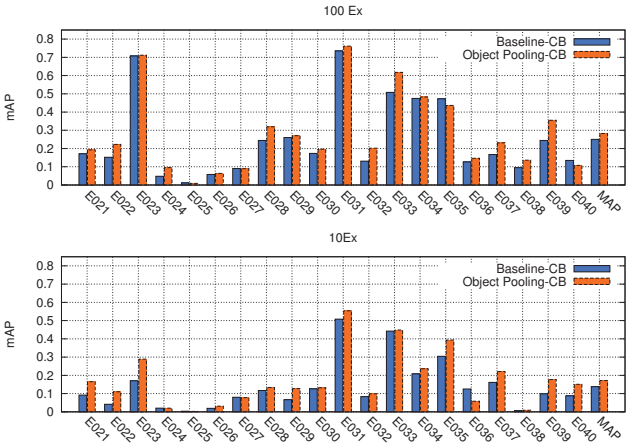

Fig. 4 MED14_Test 100/10Ex per event performance comparision with *concept-bank* feature (mAP in percentage)

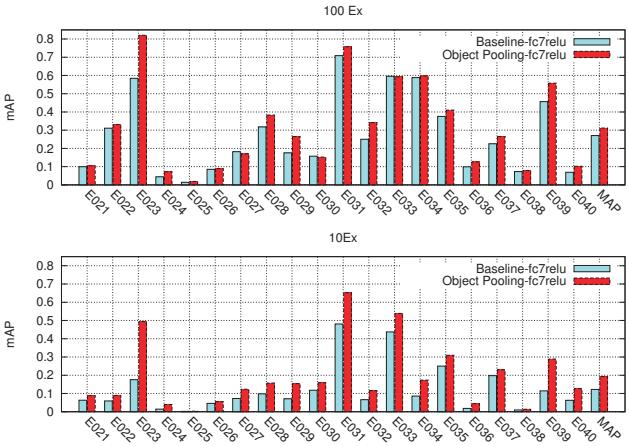

Fig. 5 MED14_Test 100/10Ex per event performance comparision with *latent-concept* descriptors (mAP in percentage)

5.4 Evaluation Details

In all the experiments, we apply homogeneous kernel mapping with VLFeat toolbox[32] and linear SVM with LIBSVM toolkit[33]. We conduct extensive experiments on two standard training conditions: in 100Ex, 100 positive exemplars are provided for each event; in 10Ex, 10 positive exemplars are provided for each event.

We uniformly sample one frame every two seconds from video and apply object pooling on each frame. For temporal pooling of frame-level feature, both *max* and *mean* operator are experimented. In the 100Ex condition, we utilize 5-fold cross-validation to select the parameter of regularization coefficient C in linear SVM. In the 10Ex condition, we set the same C as 100Ex. Average precision (AP) is used as evaluation metric.

6. Result Discussion

6.1 Results for MED

(1) Performance of object pooling

For 100Ex and 10Ex, we compare object pooling against a baseline where concept detection is conducted

directly on frame level (i.e., primary object). We conduct experiments on *concept-bank* feature (1,843 Concepts) and *latent-concept* feature (i.e., 4,096 dimensional fc7_relu). To generate video-level representation, we utilize max operator as temporal pooling strategy.

As shown in Fig 4, we observe similar patterns that object pooling outperforms baseline for most events under both 100Ex and 10Ex conditions. Specifically, with *concept-bank* feature, object pooling outperforms baseline by a relative mAP improvement of 12.8% (0.250 to 0.282) under 100Ex and 24.6% (0.138 to 0.172) under 10Ex.

With *latent-concept* descriptors (See Figure 5), object pooing also outperforms baseline by a relative mAP of 15.5% (0.27 to 0.312) and 58.2% (0.122 to 0.193), indicating that object pooling brings larger improvements especially with less training data (10Ex VS 100Ex) and has generalization ability to even latent-concept features.

(2) Impacts of Different Concept Sets

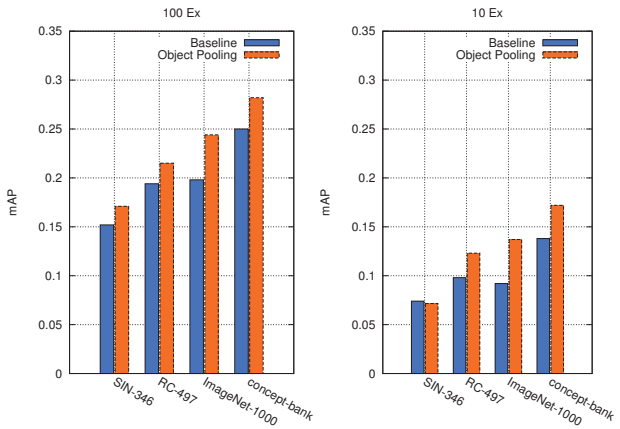

Fig. 6 MED14_Test 100/10Ex performance comparision with different concept sets (mAP)

We take subsets of the *concept-bank* feature under MED14_Test 100Ex and 10Ex as examples to see the impacts of different concept sets, and verify the robustness of object pooling.

Since concept-bank is composed of three subsets: SIN-346 (346-D), RC-497 (497-D), ImageNet-1000 (1,000-D), we compare object pooling with baseline on each subset of concepts. As shown in Figure 6, though different concept sets are used, we can see clearly that object pooling significantly outperforms baseline in most cases, indicating the robustness of object pooling. Additionally, performance of object pooling consistently improves with respect to the amount of concepts used, indicating the potential of object pooling with even larger-scale concept-bank.

EK100	mean		max	
	Baseline	OP	Baseline	OP
E021	**0.155**	0.104	0.137	**0.138**
E022	0.061	**0.084**	0.159	**0.175**
E023	0.452	**0.563**	0.413	**0.555**
E024	0.023	**0.038**	0.033	**0.059**
E025	0.012	**0.022**	**0.006**	0.005
E026	0.047	**0.048**	0.054	**0.056**
E027	**0.076**	0.065	0.066	**0.069**
E028	0.167	**0.292**	0.193	**0.264**
E029	0.144	**0.244**	0.058	**0.196**
E030	0.104	**0.138**	**0.172**	0.160
E031	0.611	**0.670**	0.729	**0.740**
E032	0.068	**0.122**	0.092	**0.220**
E033	0.359	**0.400**	0.451	**0.499**
E034	**0.390**	0.352	0.436	**0.461**
E035	0.380	**0.405**	0.358	**0.396**
E036	0.080	0.080	0.129	**0.150**
E037	0.168	**0.294**	0.158	**0.234**
E038	**0.136**	0.116	0.080	**0.101**
E039	0.329	**0.399**	0.177	**0.309**
E040	**0.047**	0.043	0.070	**0.093**
mAP	0.190	**0.224**	0.199	**0.244**

Table 2 MED14_Test 100Ex per event performance comparision with *mean*/*max* pooling in temporal domain(mAP)

EK10	mean		max	
	Baseline	OP	Baseline	OP
E021	0.065	**0.089**	0.052	**0.094**
E022	0.017	**0.025**	0.028	**0.091**
E023	**0.156**	0.088	0.052	**0.097**
E024	**0.018**	0.014	0.004	**0.009**
E025	0.004	0.004	0.002	0.002
E026	**0.027**	0.026	0.015	**0.034**
E027	0.053	**0.071**	0.049	**0.072**
E028	0.053	**0.140**	0.075	**0.123**
E029	0.063	**0.171**	0.014	**0.096**
E030	0.014	**0.018**	0.026	**0.080**
E031	0.447	**0.520**	0.358	**0.499**
E032	0.056	**0.137**	**0.094**	0.091
E033	0.254	**0.282**	0.276	**0.361**
E034	0.067	**0.084**	0.155	**0.183**
E035	0.278	**0.377**	0.250	**0.340**
E036	0.018	**0.030**	**0.057**	0.051
E037	0.087	**0.205**	0.217	**0.278**
E038	0.008	**0.009**	0.008	**0.009**
E039	0.083	**0.159**	0.077	**0.140**
E040	**0.045**	0.037	0.041	**0.093**
mAP	0.091	**0.124**	0.093	**0.137**

Table 3 MED14_Test 10Ex per event performance comparision with *mean*/*max* pooling in temporal domain(mAP)

(3) Impact of Temporal Pooling

We take ImageNet-1000 concept feature under MED14_Test 100Ex and 10Ex as examples to see the impacts of different temporal pooling methods (i.e., *max*/*mean* operator).

As is shown in Table 2, we observe that, under 100Ex, object pooling using *mean* operator in temporal pooling obtains better AP for 14 out 20 events bringing a relative mAP improvements of 17.4% (0.190 to 0.223). Similarly, *max* operator in object pooling obtains better AP for 18 out of 20 events bringing a relative mAP improvements of 23.2% (0.198 to 0.244). Similar observations can be found in Table 3 for 10Ex. To conclude, our proposed object pooling is robust to different pooling strategies, and consistently brings improvements for most of the multimedia events compared with baseline, indicating the effectiveness of regional objects. Additionally, compared with temporal pooling by operator *mean*, operator *max* contributes larger improvements. This is probably because responses of regional objects

	10Ex		100Ex	
	SP	OP	SP	OP
E021	0.085	**0.094**	0.120	**0.138**
E022	0.016	**0.091**	0.111	**0.175**
E023	0.048	**0.097**	**0.592**	0.555
E024	0.006	**0.009**	0.016	**0.059**
E025	**0.003**	0.002	**0.017**	0.005
E026	0.023	**0.034**	**0.068**	0.056
E027	**0.076**	0.072	**0.087**	0.069
E028	0.089	**0.123**	0.225	**0.264**
E029	0.096	0.096	0.144	**0.196**
E030	0.037	**0.080**	**0.176**	0.160
E031	**0.546**	0.499	0.703	**0.740**
E032	**0.149**	0.091	0.117	**0.220**
E033	0.163	**0.361**	**0.518**	0.499
E034	0.070	**0.183**	0.389	**0.461**
E035	0.316	**0.340**	**0.427**	0.396
E036	0.027	**0.051**	0.063	**0.150**
E037	0.211	**0.278**	**0.236**	0.234
E038	0.009	0.009	0.100	**0.101**
E039	0.108	**0.140**	**0.395**	0.309
E040	0.033	**0.093**	0.056	**0.093**
mAP	0.106	**0.137**	0.228	**0.244**

Table 4 MED14_Test 100Ex per event performance comparision (mAP)

are downgraded by mean pooling in the process of generating video-level representation, reducing the effects of regional objects.

(4) Object pooling and Spatial pyramid

To verify effectiveness of regional objects against rigidly divided pyramid, we also compare performances of object pooling against spatial pyramid (SP)[3]. By SP, a frame is partitioned into a pyramid of 1×1 and 2×2 regions, for which each region is represented by an object histogram. SP temporally pools the histograms by *max* operator, resulting in five video-level histograms, which are finally early fused and fed into SVM for classifier learning. As shown in Table 4, object pooling still outperforms SP by a relative mAP improvement of 18.1% (0.108 to 0.137) for 10Ex and a relative mAP improvement of 7% (0.228 to 0.244) for 100Ex. Note that our proposed method adopts a more compact feature (2,000 dimensions) than SP (5,000 dimensions).

6.2 Results for Evidence Localization

Given an event query, we retrieve relevant videos from testing video set (i.e., MED14_Test) by MED system. Then, we identify the importance of each object by calculating its contribution to the decision function (4). We treat the top few objects which contribute most to decision function (4) as evidential objects. Finally, with the help of object pooling, we can easily identify *spatial* and *temporal* locations of evidentiall objects in these videos. We show some of the identified evidential objects in Figure 7.

We observe that for event "bike trick", the evidential objects, such as "trike", "bike", "moped" and etc, are all relevant objects. For most of them, our system is able to locate objects with their spatial regions. However, we also observe a special case: when primary object is selected as evidences, the whole frame is recommended as evidence, which is reasonable. For other

Fig. 7 Examples of evidences located for videos of different events. The film strip highlights temporal evidences, and the bounding box shows the spatial position of an evidence concept. The key concept associated with a bounding box is also given.

events, (e.g., "dog show", and "beekeeping"), there are few irrelevant objects involved, such as "green table" for event "dog show" and "loaf" for event "beekeeping". The reason lies in that "green table" is visually similar to "carpet" which is evidential objects in "dog show" and "loaf" is visually similar to "honeycomb" which is evidential objects for "Beekeeping". To conclude, object pooling are able to locate evidential objects in both temporal and spatial domain with a reasonable performance.

7. Conclusions

We have presented object pooling for dynamic localization of spatio-temporal evidences. Experimental findings suggest that the approach is effective for event detection in web videos, especially when very few positive training examples are available. When applying for event evidence localization, object pooling also demonstrates potential in enabling a more quick and accurate way of judging video relevancy.

Acknowledgment

The work described in this paper was supported by two grants from the Research Grants Council of the Hong Kong Special Administrative Region, China (CityU 120213) and National Natural Science Foundation of China (No. 61272290).

References

1) P. Over, G. Awad, M. Michel, J. Fiscus, G. Sanders, W. Krarij, A. Smeaton and G. Queenot: "TRECVID 2014-an overview of the goals, tasks, data, evaluation mechanisms and metrics", 2014 TREC Video Retrieval (2014)
2) S. Bhattacharya, F.-X. Yu, and S.-F. Chang: "Minimally needed

evidence for complex event recognition in unconstrained videos", ACM International Conference on Multimedia Retrieval (2014)
3) S. Lazebnik, C. Shmid, and J. Ponce: "Byond bags of features: Spatial pyramid matching for recognizing natural scene categories", IEEE Conference on Computer Vision and Pattern Recognition (2006)
4) W. Li, Q. Yu, A. Divakaran and N. Vasconcelos: "Dynamic pooling for complex event recognition", IEEE International Conference on Computer Vision, pp. 2728-2735 (2013)
5) Z.-W. Xu, Y. Yang and A. Hauptmann: " A Discriminative CNN Video Representation for Event Detection", IEEE Conference on Computer Vision and Pattern Recognition (2015)
6) P. Mettes, J. Gemert, S. Cappallo, T. Mensink and C. Snoek: "Bag-of-Fragments: Selecting and encoding video fragments for event detection and recounting", ACM International Conference on Multimedia Retrieval (2015)
7) K. Lai, F.-X. Yu, M.-S. Chen and S.-F.Chang: "Video event detection by inferring temporal instance labels", IEEE Conference on Computer Vision and Pattern Recognition, pp. 2251-2258 (2014)
8) K. Lai, D. Liu, M.-S. Chen and S.-F.Chang: "Recognizing complex events in videos by learning key static-dynamic evidences", Computer Vision-ECCV, pp. 675-688 (2014)
9) Y. Yang, Z. Ma A. Hauptmann and N. Sebe: "Feature selection for multimedia analysis by sharing information among multiple tasks", IEEE Trans. on Multimedia, pp. 661-669 (2013)
10) M. Mazloom, E. Gavves and C. Snoek: "Conceptlets: Selective semantics for classifying video events", IEEE Trans. on Multimedia, 16(8), pp. 2214-2228 (2014)
11) A. Habibian and C. Snoek: "Recommendations for recognizing video events by concept vocabularies", Computer Vision and Image Understanding, pp. 110-122 (2014)
12) C. Gan, N. Wang, Y. Yang, D.-Y. Yeung and A. Hauptmann: "Devnet: A Deep Event Network for Multimedia Event Detection and Evidence Recounting", IEEE Conference on Computer Vision and Pattern Recognition (2015)
13) W. Wang, E. Yeh: "ISOMER:Informative segment observations for multimedia event recounting", ACM International Conference on Multimedia Retrieval (2014)
14) C. Sun, R. Nevatia: "DISCOVER: Discovering important segments for classification of video events and recounting", IEEE Conference on Computer Vision and Pattern Recognition, pp. 2569-2576 (2014)
15) Q. Yu, J. Liu, H. Cheng, A. Divakaran and H. Sawhney: "Multimedia event recounting with concept based representation", ACM International Conference on Multimedia , pp. 1073-1076 (2012)
16) L.-J. Li, H. Su, F.-F. Li, and E.-P. Xing: "Object bank: A high-level image representation for scene classification & semantic feature sparsification", Advances in Neural Information Processing Systems, pp. 1378-1386 (2010)
17) Y. Tian, R. Sujthankar and M. Shah: "Spatiao temporal deformable part models for action detection", IEEE Conference on Computer Vision and Pattern Recognition, pp. 2642-2649 (2013)
18) A. Krizhevsky, I. Sutskever and G.E. Hinton: "ImageNet classification with deep convolutional neural networks", Advances in Neural Information Processing Systems, pp. 1097-1105 (2012)
19) R. Girshick, J. Donahue, T. Darrell and J. Malik: "Rich feature hierachies for accurate object detection and semantic segmentation", IEEE Conference on Computer Vision and Pattern Recognition, pp. 580-587 (2014)
20) P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, D. Fergus and Y. LeCun: "OverFeat: Integrated recognition, localization and detection using convolutional networks", International Conference on Learning Representations (2014)
21) J. Liu, Q. Yu, O. Javed, S. Ali, A. Tamrakar, A. Divakaran, H. Cheng and H. Sawhney : "Video event recognition using concept attributes", IEEE Workshop on Applications of Computer Vision, pp. 339-346 (2013)
22) M. Mazloom, A. Habibian and C. Snoek: "Querying for video events by semantics signatures from few examples", ACM International Conference on Multimedia, pp. 609-612 (2013)
23) E. Can and R. Manmatha: "Modeling concept dependencies for event detection", ACM International Conference on Multimedia Retrieval, p. 289 (2014)
24) A. Habibian, T. Mensink and C. Snoek: "Videostory: A new multimedia embedding for few-example recognition and translation of events", ACM International Conference on Multimedia, pp. 17-26 (2014)
25) J. Uijlings, K. Vandesande, T. Gevers, and A. Smeulders: "Selective search for object recognition", International Journal of Computer Vision, pp. 154-171 (2013)
26) A. Bogdan, T. Deselaers, and V. Ferrari "Measuring the objectness of image windows IEEE Trans. Pattern Analysis and Machine Intelligence, pp. 2189-2202 (2012)
27) M.-M. Cheng, Z. Zhang, W.-Y. Lin and P. Torr: "BING: Binarized normed gradients for objectness estimation at 300fps", IEEE Conference on Computer Vision and Pattern Recognition pp. 3286-3293 (2014)
28) A. Vedaldi, and A. Zisserman: "Efficient additive kernels via explict feature maps", IEEE Trans. IEEE Trans. on Pattern Analysis and Machine Intelligence, pp. 480-492 (2012)
29) Y.-Q Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama and T. Darrell: "Caffe: convolutional architecture for fast feature embedding", ACM International Conference on Multimedia, pp. 675-678 (2014)
30) J. Deng, W. Dong, R. Socher, L.-J Li, K. Li and F.-F Li: "ImageNet: A large-scale hierachical image database", IEEE Conference on Computer Vision and Pattern Recognition, pp. 248-255 (2009)
31) P. Natarajan, S. Wu, F. Luisier, X. Zhuang, and M. Tickoo: "BBN VISER TRECVID 2013 multimedia event detection and multimedia event recounting systems", 2013 TREC Video Retrieval Workshop, (2013)
32) A. Vedaldi and B. Fulkerson: " VLFeat: An Open and portable library of computer vision algorithms", ACM International Conference on Multimedia, pp, 1469-1472 (2010)
33) R. Fan, K. Chang, C.-J. Hsieh, X.-R. Wang, C.-J. Lin: "LIBLINEAR: A Library for Large Linear Classification", 2008 Journal of Machine Learning Research, pp. 1871-1874 (2008)
34) K. Simonyan and A. Zisserman: "Very deep convolutional networks for large-scale image recognition", arXiv preprint 2014 arXiv:1409.1556 (2014)

Hao Zhang received the B.Sc. degree from Nanjing University, Nanjing, China, in 2012, the M.Sc. degree from the Chinese University of Hong Kong, Hong Kong, China, in 2013. He is currently working towards the Ph.D. degree in computer science at the City University of Hong Kong, Hong Kong.
He is currently with VIREO Group, City University of Hong Kong. His research interest lies in multimedia content analysis, including Semantical Concept Indexing and Multimedia Event Detection

Chong-Wah Ngo received the B.Sc. and M.Sc. degrees in computer engineering from Nanyang Technological University, Singapore, and the Ph.D. in computer science from the Hong Kong University of Science and Technology (HKUST), Hong Kong.
He is currently a Professor with the department of Computer Science, City University of Hong Kong, Hong Kong. Before joining the City University of Hong Kong, he was a Postdoctoral Scholar with the Beckman Institute, University of Illinois at Urbana-Champaign (UIUC), Urbanna, IL, USA. He was also a Visiting Researcher with Microsoft Research Asia, Beijing,China. His research interests include large-scale multimedia information retrieval, video computing, multimedia mining, and visualization.
Prof. Ngo was the Associate Editor of the IEEE TRANSACTIONS ON MULTIMEDIA (2011-2014). He was the Conference Co-Chair of the ACM International Conference on Multimedia Retrieval 2015 and the Pacific Rim Conference on Multimedia 2014. He also served as Program Co-Chair of ACM Multimedia Modeling 2012 and ICMR 2012. He was the Chairman of ACM (Hong Kong Chapter) from 2008 to 2009