Modeling Concept Dependencies for Event Detection

Ethem F. Can
Center for Intelligent Information Retrieval (CIIR)
School of Computer Science
UMass Amherst, MA, 01002
efcan@cs.umass.edu

R. Manmatha
Center for Intelligent Information Retrieval (CIIR)
School of Computer Science
UMass Amherst, MA, 01002
manmatha@cs.umass.edu

ABSTRACT

Event detection is a recent and challenging task. The aim is to retrieve the relevant videos given an event description. A set of training examples associated with the events are generally provided as well, since retrieving relevant videos from textual queries solely is not feasible. Early attempts of event detection are based on low-level features. High level features such as concepts for event detection have been introduced as an alternative to low-level features since high-level features provide semantically richer information. In this work, we focus on object-based concepts and exploit their dependencies using a Markov Random Field (MRF) based model for event detection. This enables us to model likelihood of concepts, either pairwise or individually, present in the videos. Here, we propose a method incorporating the strengths of concepts and MRF based model for event detection task. We evaluate our models on an Multimedia Event Detection (MED) dataset from NIST's 2011 TRECVID Multimedia, which consists of approximately 45,000 unconstrained videos. This type of work is beneficial from several respects. First, we focus on the task of concept-based event detection using a very large number of unconstrained Youtube videos. Second, we introduce the application of MRF's for the event detection purpose, which can further be enhanced incorporating other features or temporal information. At last but not means least, we exploit the occurrence and co-occurrence of object based concepts for event detection that enables us to reveal interactions of such concepts in the video level. Experimental results show that revealing these interactions provide promising event detection results.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Retrieval models; I.2.10 [**Computing Methodologies**]: Artificial Intelligence—*Vision and Scene Understanding*

http://dx.doi.org/10.1145/2578726.2578763
ICMR '14, Apr 01-04 2014, Glasgow, United Kingdom

General Terms

Algorithms, Experimentation

Keywords

Concept based event detection, Concept dependencies, MRF model, Multimedia event detection, Object based concepts

1. INTRODUCTION

In recent years there has been considerable interest in finding ways to search large amount of data generated in the form of videos. A recent report indicates that there were more than 50 billion views by U.S. Internet audience on December 2013 [3]. Searching video clips online becomes crucial since 50 billion views should be preceded with a huge number of searches.

The specific problem tackled in this paper is event detection in complex videos where we are given a query with a set of example videos and our task is to rank the videos in terms of relevance to the given query. A query could include text, audio and other modalities but in this paper we focus on the visual aspects. This procedure produces a ranked list of videos where the videos relevant to the queried event are located higher in the list than the non-relevant ones.

Event detection is challenging since the large number of videos and frames mean that data processing is often an issue. Further, many videos on sites such as YouTube are unconstrained in how they are recorded (often by amateurs) and suffer from a number of issues including low-resolution, high degrees of motion blur, and camera motion. Even though event detection and action recognition suffer from the same issues, event detection is different from action recognition where a fair amount of work has been done [17, 21]. The main difference is that, events can be composed of one or more actions. For example, events such as "Birthday Party" may not necessarily involve any actions while other events such as "Parkour" may involve multiple atomic actions. Further, videos may vary widely in length and the event may form only a small part of the video. Therefore, event detection can be considered more difficult than action recognition.

The best results on event detection have involved running classifiers on low level features [20, 16]. One approach to using concepts for video retrieval (e.g the LSSCOM effort [15]) involves detecting specific concepts in videos and running text queries against those concept outputs. There are two limitations to most of the concept based approaches. First, events are not simple enough to be revealed using only

the occurrences of the concepts. Second, most of these concept detector creations are based on the given dataset. In other words, concept detectors are trained on the event detection datasets. Here as an alternative to training concept detectors on the event detection sets, we learn our object-based concept detectors on the images. Then we make use of such detectors to measure the likelihood of the concepts in the videos. The first issue is addressed by exploiting co-occurrences of the concepts in addition to their occurrences. We make use of an MRF based model to exploit dependencies of the object-based concepts. Here, we focus on two different dependency settings; full independence and spatial dependence. We aim to use occurrences and co-occurrences of the object-based concepts to obtain a higher level of representation for the unconstrained videos for event detection. Exploiting such occurrences enables us to reveal the interactions of object-based concepts in the videos.

For example, the occurrence of tire and car concepts play an important role for discriminating the events involved with a vehicle such as "driving a car"' and "visiting a tire shop"; however, the co-occurrence of these concepts is more discriminative for the events such as "changing a tire". The co-occurrence of car and tire concepts may imply changing a car tire; however, individual occurrences of these concepts may imply different semantics than changing a tire.

In Figure 1 we summarize our method. We start with getting responses of object-based concept detectors for video frames. Then we feed those responses to our MRF based model. After exploiting dependencies of the concepts, we rank videos using the outputs of our models. We create our object-based concepts using a number of static images from the ImageNet [6].

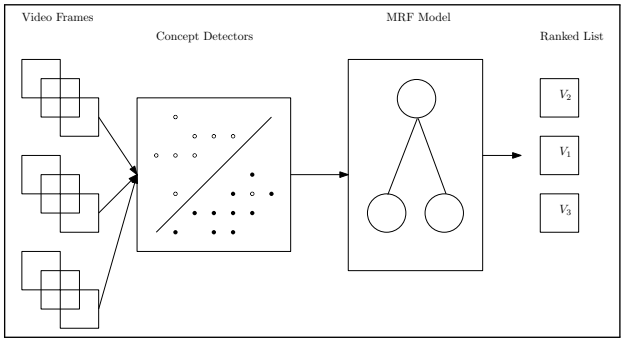

Figure 1: Illustration of the proposed approach.

We empirically show on the 2011 TRECVID multimedia event detection (MED) task that exploiting concept dependencies using an MRF based retrieval model performs as good as the best low level features and outperforms them for many events. Even though our aim is detection rather than recognition, we compare our models with an event recognition study where state-of-the-art object detectors such as ObjectBank/DetectionBank are used and show that our models outperform state-of-the-art detectors.

The rest of the paper is organized as follows; we first provide a summary of the previous work on event detection. It is followed by the problem formulation that consists of a retrieval model and concept detection. We then provide information about the experimental settings. Afterwards experimental results and discussion are presented. We finally conclude the paper.

2. RELATED WORK

In this section, we briefly summarize the related work in the event detection literature. Ballan et al. [2] present a method to introduce temporal information for video event detection with a BoW (bag-of-words) approach. Zhou et al. [24] study video event detection by encoding a video with a set of bag of SIFT feature vectors and describe the distribution with a Gaussian Mixture Model (GMM). Jiang et al. [8] provide a summary of experiments for TRECVID MED 2010. They employ low-level features such as SIFT and spatial-temporal interest points (STIP) to capture time-space volumes. They also compute a histogram of oriented gradients (HOG) and a histogram of optical flow (HOF) in 3D video patches. The audio is represented using mel-frequency cepstrum (MFC). In addition to low-level features, they also study the creation of concept detectors. According to their results, the fusion of STIP, SIFT and MFC provides better results than using individual descriptors. Besides, they report that STIP is slightly better than SIFT in terms of average precision (AP). Tamrakar et al. [20] also show that by fusing a large number of low level detectors they can get state of the art performance for event detection on the NIST MED'11 task. They also introduce trajectory based features called DTF-HOG and DTF-MBH with the former being the best performing feature in their set. Natarajan et al. [16] do event detection with low level features as well as other modalities but demonstrate that low level features alone give good performance.

Snoek et al. [19] focus on indexing a number of semantic concepts. Ma et al. [12] address two basic problems in event detection; detecting more generic and complicated events which is the major goal of TRECVID MED task, and detection using fewer examples. Lan et al. [23] introduce double fusion –combination of early and late fusion– as a competitor to early and late fusion. Althoff et al. [1], Izidinia and Shah [7], and Habibian et al. [5] focus on event recognition rather then event detection (i.e., the main focus is recognition but not ranking) Yang and Shah [22] discover data-driven concepts from multi-modality signals (audio, scene and motion) to describe high level semantics of videos. The ObjectBank representation [10] creates a histogram of objects by running multiple object detectors - these are not a per image classifier but actually try to return the location of the object. This was originally used to classify an image into scenes. They apply their technique to events in static images. A variation of ObjectBank called DetectionBank [1] has been applied to event recognition. ObjectBank based techniques are computationally expensive since each object detector must be run over every frame in the video. Besides, DetectionBank is applied for the task of event recognition where the information from other events are used as in the case of action recognition. We also note the ActionBank [18] work for the related task of action recognition where videos are used to train a set of temporal action detectors and then the distribution over these actions is used to recognize them. Again this is very computationally expensive and time consuming. Liu et al. [11] and Mazloom et al. [13] also focus on concepts for event detection. Most of the work above either focus on event recognition, not event detection like the proposed approach. Event recognition is a different task, where a number of events are involved in the same model, than

event detection in which we create binary ranking models per event. In other words, event recognition aims to recognize events of a video, whereas event detection aims to rank videos with respect to their relevance to an event query. From the perspective of studies that focus on event detection, they do not exploit dependencies of the concepts for event detection as we do.

3. PROPOSED APPROACH

In our work we focus on a number of object-based concepts and use these concepts for given a test event. We exploit the dependencies of such concepts for a better retrieval. In the next section, we explain the MRF based retrieval model that is used to model the dependencies of concepts.

3.1 Retrieval Model

Here, we describe a Markov random field (MRF) approach to model concept dependencies. An MRF is commonly used in the machine learning domain and is an undirected graphical model. It models the joint distributions. A Markov random field is constructed using a graph G. The nodes represent the random variables and edges define the dependencies [14]. In our work, we assume G consists of concepts $C = \{c_1, c_2, ..., c_n\}$ and a video node V. The joint distribution over the random variables in G is defined as follows;

$$P_\Lambda(C, V) = \frac{1}{Z_\Lambda} \prod_{l \in L(G)} \psi(l; \Lambda) \tag{1}$$

where L is the set of cliques in G and ψ is a non-negative potential function over a clique parameterized by Λ. $Z_\Lambda = \sum_{C,V} \prod_{l \in L(G)} \psi(l; \Lambda)$ is a normalizer and is very costly to compute [14, 4]. Removing the normalizer does not change the final ranking of the videos.

Metzler and Croft [14] used an MRF based model to exploit term dependencies, and Feng and Manmatha [4] used a similar approach for single image retrieval. We modify the approach used in these studies and the posterior for ranking can therefore be computed as follows:

$$\begin{aligned} P_\Lambda(V|C) &= \frac{P_\Lambda(C, V)}{P_\Lambda(C)} \\ &\overset{rank}{=} logP_\Lambda(C, V) - logP_\Lambda(C) \\ &\overset{rank}{=} \sum_{l \in L(G)} log\psi(l; \Lambda) \end{aligned} \tag{2}$$

Potential functions are usually parametrized as follows;

$$\psi(l; \Lambda) = e^{\lambda_l f(l)} \tag{3}$$

where $f(l)$ is a real-valued feature function over clique values and λ_l is the weight for that particular feature. Then Equation 2 becomes;

$$P_\Lambda(V|C) \overset{rank}{=} \sum_{l \in L(G)} \lambda_l f(l) \tag{4}$$

We would like to focus on two different dependency settings in our model; 1) fully independent concepts f_I, and 2) spatially dependent concepts f_{S_d}. Expanded with two different dependency settings, Equation 4 becomes:

$$P_\Lambda(V|C) = \sum_{l \in I} \lambda_I f_I(l) + \sum_{l \in S_d} \lambda_{S_d} f_{S_d}(l) \tag{5}$$

where I is defined to be the cliques containing a concept and a video V, S_d is the set of cliques consisting of a video node V and two concepts occurring spatially together. Even though we formulate our dependency models together here, we also provide results of the models individually. Below we explain the potential functions that we use in our work.

3.1.1 Potential Functions (ψ)

Full Independence Model (ψ_I)

The potential function for the fully independent model can be defined over 2-cliques where there is an edge between a concept c_i and the video V. This potential function measures how likely a concept c_i is in the video. We define our first potential function in the following way;

$$\psi_I(l) = \lambda_I f_I(l) \tag{6}$$

where the feature function is defined over 2-cliques of concepts c_i where $i = 1, ..., n$ and a video V. A video V consists of a number of video frames $V = \{v_1, v_2, ...v_t\}$. Therefore, the function becomes (when we substitute V with a number of video frames $\{v_1, v_2, ...v_t\}$);

$$\begin{aligned} \psi_I(l) &= \lambda_I f_I(l) \\ &= \lambda_I \omega_l \frac{\log \sum_t \delta(c_i, v_t)}{\max_i \{\log \sum_t \delta(c_i, v_t)\}} \end{aligned} \tag{7}$$

where $\delta(c_i, v_t)$ is the response of the concept c_i obtained from object detectors at video frame v_t. We explain the concept detection and responses in the next section in detail. Here ω_l is defined to be a co-efficient associated with a concept c_i in a clique l. When we substitute the function into Equation 5, it becomes;

$$\begin{aligned} &= \sum_{l \in I} \lambda_I f_I(l) + \sum_{l \in S_d} \lambda_{S_d} f_{S_d}(l) \\ &= \sum_{l \in I} \lambda_I \omega_l \frac{\log \sum_t \delta(c_i, v_t)}{\max_i \{\log \sum_t \delta(c_i, v_t)\}} + \sum_{l \in S_d} \lambda_{S_d} f_{S_d}(l) \end{aligned} \tag{8}$$

In Figure 2, we illustrate a fully independent model for an example in which there are two video frames $V = \{v_0, v_1\}$ and three concepts c_1, c_2, c_3. The cliques for the given example are (v_0, c_1), (v_0, c_2), (v_0, c_3), (v_1, c_1), (v_1, c_2), and (v_1, c_3).

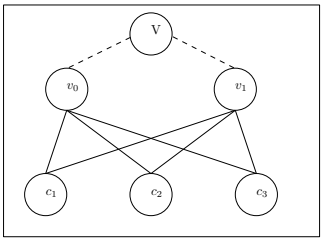

Figure 2: Illustration of fully independent model over a video $V = \{v_0, v_1\}$ and concepts c_1, c_2, c_3.

Spatial Dependence Model (ψ_{S_d})

In addition to fully independent model, we also exploit the spatial dependencies between concept pairs. Here, we consider pairs of concepts where they are spatially dependent. A spatial dependence is defined if two concepts occur together at the same video frame v_t. In this model, the dependencies are formed between a video frame and a pair of concepts. The spatial dependence model is illustrated for an example where a video V consists of two video frames v_0, v_1 and there are three concepts c_1, c_2, c_3 in Figure 3. In the example, the cliques are (v_0, c_1, c_2), (v_0, c_1, c_3), (v_0, c_2, c_3), (v_1, c_1, c_2), (v_1, c_1, c_3), and (v_1, c_2, c_3). In Figure 3, there are edges between concepts; whereas, these edges are removed in fully independent model (see Figure 2).

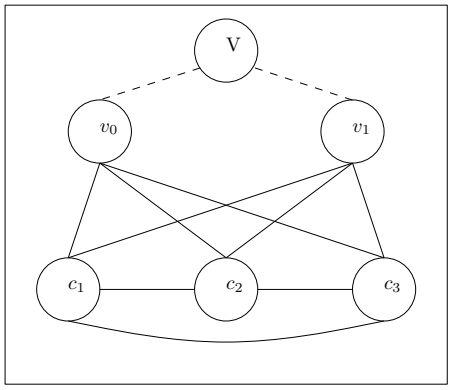

Figure 3: Illustration of spatial dependence model over video frames $V = \{v_0, v_1\}$, and concepts c_1, c_2, c_3.

For the spatial dependence model we define the potential function considering the video frames in the following way;

$$\begin{aligned} \psi_{S_d}(l) &= \lambda_{S_d} f_{S_d}(l) \\ &= \lambda_{S_d} \omega_l \frac{\log \sum_t \delta(c_i, c_j, v_t)}{\max_i \{\log \sum_t \ \delta(c_i, c_j, v_t)\}} \end{aligned} \tag{9}$$

where $\delta(c_i, c_j, v_t)$ is the response of concepts c_i and c_j at video frame v_t. Similar to the fully independent model, here ω_l is defined to be a co-efficient associated with a concept c_i and c_j in a clique l. When we substitute the functions into Equation 5, it becomes;

$$\begin{aligned} &= \sum_{l \in I} \lambda_I f_I(l) + \sum_{l \in S_d} \lambda_{S_d} f_{S_d}(l) \\ &= \sum_{l \in I} \lambda_I f_I(l) + \sum_{l \in S_d} \lambda_{S_d} \omega_l \frac{\log \sum_t \delta(c_i, c_j, v_t)}{\max_i \{\log \sum_t \ \delta(c_i, c_j, v_t)\}} \end{aligned} \tag{10}$$

3.2 Concept Detection

In this work, we create our concept detectors on top of the images derived from ImageNet rather than creating them using the training videos. In this way, for different sets of training and test videos we do not have to re-create our concept models; therefore, they are independent of the training and test videos.

We focus on 1,000 object-based concepts and 100 images from ImageNet [6] for each concept. We extract dense SIFT (DSIFT) features at multiple scales (100%, 50%, and 25%) with a step size of 10 pixels and the width and height of the images are bounded by 300 pixels). We quantize the raw DSIFT descriptors using a vocabulary of size 1,000 and obtain the bag-of-word (BoW) representation for each image. Having quantized raw descriptors and obtained BoW representations, we follow a simple but effective method to create concept detectors. We use a multi-class SVM and the BoW representation for each image to create a multi-concept model. In order to get the concept responses, we first split videos into frames by sampling 3 frames/second. We create a BoW representation for each frame using DSIFT descriptors. We then test each video frame against the model of the concept detectors. Since we have 1,000 concepts for each frame we have 1,000 concept outputs (Φ_i output of concept c_i). The outputs measure how much a specific concept is likely to be in a video frame. The larger the output of a concept, the higher the chance that the concept is in that video frame. We make use of these outputs to compute δ functions. Note that concept detectors are independent of event detection. and we do not use any information from the videos of event detection datasets.

Given that our aim is to use occurrences and co-occurrences of the object-based concepts to obtain a higher level representation of videos, while using the knowledge from image level to a higher (i.e. video) level we only choose the concepts that may be more representative than the others. We select the k-highest responses among the others instead of using all the concepts. We try to move to a higher (video) level while carrying as much knowledge as possible from the lower (image) level, considering the computation cost. Limiting the number of concepts to be considered in the MRF framework improves the scalability and reduces the data requirements. At the same time, it provides a sparser representation and discards the concepts where the probability of occurrences are too low to be representative. Keeping these motivations in mind, we define $\delta(c_i, v_t)$ as follows;

$$\delta(c_i, v_t) = \begin{cases} 1, & \text{if } \Phi_i \geq \Phi' \\ 0, & \text{otherwise} \end{cases}$$

where Φ_i is the output of the concept c_i at video frame v_t and Φ' is the k^{th} maximum response among the concept responses for video frame v_t. Similarly the function $\delta(c_i, c_j, v_t)$ is defined to be as follows;

$$\delta(c_i, c_j, v_t) = \begin{cases} 1, & \text{if } \Phi_i \geq \Phi' \text{ and } \Phi_j \geq \Phi' \\ 0, & \text{otherwise} \end{cases}$$

The δ function enables us to focus on the concepts which have a relatively higher chance to be in a video frame. In our work, we focus on k concepts for a video frame, where those concepts have the k maximum responses among all of the concepts. We search for the best k value on a validation set and explain this process in Experimental Setup section.

4. EXPERIMENTAL SETUP

In order to test the performance of our models, we focus on NIST's TRECVID MED 2011 evaluation set which uses ten events. We have chosen this dataset since most of the previous work in event detection focused on many different settings, and this dataset was the largest publicly available dataset published so far. We use three collections; the

event kit (EC) and a transparent development kit (DevT) for training and the DevO set for testing. These events are; *(E006) Birthday party, (E007) Changing a vehicle tire, (E008) Flash mob gathering, (E009) Getting a vehicle unstuck, (E10) Grooming an animal, (E011) Making a sandwich, (E012) Parade, (E013) Parkour, (E014) Repairing an appliance, (E015) Working on a sewing project.* The EC collection has 2,062 videos each of which is relevant to an event. The DevT collection has near misses for all events. For an event a near miss is a video where the event is not performed in the required way. For example, for the event "Landing a fish", a fishing video where no fish is landed is not counted as relevant (these judgments are NIST's not ours). The total number of videos is a little bit above 10,000 in DevT. The DevO test set has 32,061 videos. Videos may vary in length from 6 seconds to 3 hours.

In order to evaluate the experiments we calculate missed detection (MD) & false alarm (FA) rates - specifically MD at FA=5%. Even though NIST moved to MAP scores, we provide our numbers in MD scores because only MD numbers are available in this large dataset. By this way, we can directly compare our results with Tamrakar et al. [20]. False Alarm is the ratio of the number of non-positive clips when i clips are retrieved to the total number of non-positive video clips in the dataset. Missed detection is the ratio of missed positive clips when i clips are retrieved to the total number of positive video clips in the dataset for that query. A lower missed detection score means a better retrieval.

We train the parameter k on a validation set obtained by combining EC and DevT and then splitting it so that 70% is used for training and 30% for validation (we need to do this since DevT does not have any positives for the relevant events). Using a grid training procedure we obtain $k = 20$ as the optimal value. We follow the same procedure for searching the best values for the parameters λ_I and λ_{S_d}. When we search for the best co-efficients for the cliques we would like to maximize the number of video pairs -one relevant and one non relevant- so that relevant video is ranked higher than non-relevant video. Joachims [9] showed that this ranking problem can be formulated by introducing non-negative slack variables to the optimization problem. We follow the same approach to search for the best co-efficients for the models. For each event we create a model that exploits the dependencies of the concepts. Then the videos in the test bed are ranked using such models.

In our experiments we focus on two different spatial layouts. For a fully independent setting we make use of a spatial pyramid representation which is shown to be effective in computer vision. We create a pyramid with three levels (0,1,2). In Figure 4, we illustrate the spatial pyramid layout. Level 0 is the original image (one region), in level 1 each image is divided into 4 regions and in level 2 into 16 regions. For a video frame when we use a fully independent setting we have 21 regions in total since we consider level 0, level 1, and level 2 together. For the spatial-dependence case, we focus only on level 0 and level 1 together for practical reasons.

5. RESULTS AND DISCUSSION

Event recognition comparison: Although our aim in this work is event detection, we first compare our concepts with a recent attempt using state-of-the-art object detectors for event recognition [1]. In order to compare our concepts

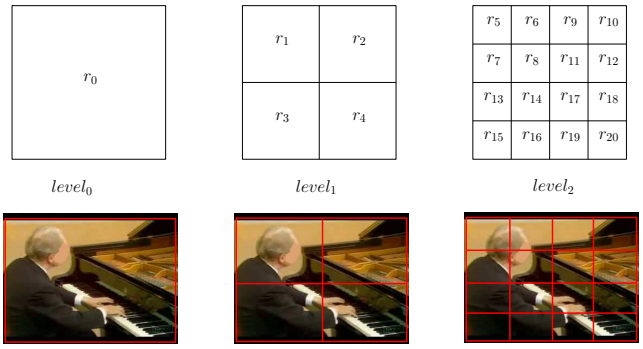

Figure 4: Illustration of the spatial pyramid layout.

with the study [1], we simulate their experimental settings since they have a different setting. They focus on a very small dataset consisting of approximately 2,000 videos and recognition is performed rather than detection. In the task of event detection, the aim is to rank the videos according to the relevance to a given test event; whereas, in event recognition it is formulated as recognizing the event of a given test video. Event recognition is a variant of action recognition. We create a multi-class SVM using our fully-independent model only. We re-formulate our clique dependencies into vector space i.e. treating each clique as a feature value. We obtain a recognition accuracy of 64.54% with our concepts; whereas, they obtain recognition accuracies of 58% with ObjectBank features, and 56% with DetectionBank features individually.

Event detection results: Table 1 shows the results on the MED11 dataset using low level features as reported by [20] as well as our results. Most columns are self-explanatory. The DTF-HoG and DTF-MBH are trajectory based features. Note that on eight of the ten events DTF-HoG is the best and on the remaining two, DTF-MBH is the best among the low-level features. The last columns show the results of our models. Although our models do not consider temporal information, they outperform the best low-level features in many cases. These results were amongst the top ones in the MED11 competition and are, therefore, a reasonable baseline. The results with low-level features were produced using SVM with an intersection kernel[20].

When we compare the fully independence and spatial dependence model, in more than half of the cases fully independence model outperforms spatial dependence model. For some of the cases spatial dependence model is better than fully independent model in terms of retrieval performance. This might be explained with the fact that for the events "E006", "E008", "E011", and "E012" co-occurrences are slightly more important than the occurrences of the individual concepts. It might be claimed that the events "birthday party", "flash mob gathering", "making a sandwich", and "parade" involves more interaction of concepts than the other events. When we consider our final model that employs both dependency settings together, we observe that exploiting such dependencies together improve the retrieval performance for all cases.

Our models provide the best MD rates for the events; "changing a vehicle tire," "flash mob gathering," "making a sandwich," "repairing an appliance," and "working on a

Table 1: Results from [20] for different features as well as of our models. The columns FI (fully-independent model; ψ_I), SD (spatial-dependence model; ψ_{S_d}), and $FI + SD$ (when we consider fully-independent and spatial-dependence model together in the MRF model) show the results of our models. Results report MD at 5% FA.

Event	GIST	SIFT	c-SIFT	mo-SIFT	STIP	DTF HoG	DTF MBH	FI	SD	FI+SD
E006	76.2	66.9	55.2	77.3	59.3	**39.0**	47.1	51.7	51.2	48.8
E007	76.1	67.3	51.3	91.2	64.6	37.2	45.1	35.4	41.1	**32.7**
E008	32.6	22.2	18.5	79.3	20.0	14.1	14.1	12.6	10.2	**9.5**
E009	71.1	49.4	44.6	78.3	43.4	**26.5**	38.6	31.3	32.9	30.4
E010	85.2	81.5	69.1	95.1	67.9	**50.6**	61.7	58.0	60.8	51.9
E011	72.3	78.8	53.3	89.8	65.0	55.5	61.3	49.6	47.7	**43.0**
E012	61.5	37.4	39.6	59.9	46.0	26.7	**18.2**	40.6	37.6	32.3
E013	67.7	55.9	45.1	90.2	22.6	19.6	**12.8**	30.4	34.7	29.7
E014	63.6	43.2	37.5	75.0	34.1	28.4	36.4	26.1	32.9	**23.2**
E015	80.5	68.3	58.5	80.5	52.4	43.9	43.9	45.1	48.7	**42.1**

Table 2: Fused results from [20] (fused) and fusion results of our models (FI+SD) with DTF-HOG and DTF-MBH ((FI+SD)+HOG+MBH). Results report MD at 5% FA.

Event	Fused	(FI+SD)+HOG+MBH
E006	29.7	**28.1**
E007	22.1	23.4
E008	7.4	**6.3**
E009	27.7	**21.5**
E010	40.7	**37.9**
E011	35.8	36.7
E012	13.9	**8.6**
E013	**8.8**	**8.8**
E014	22.7	**19.5**
E015	35.4	**31.6**
Avg.	24.4	22.2

sewing project". When we analyze the ranked list for the event "birthday party", false alarms are mostly near misses. In other words, they are videos about gatherings but they are not necessarily birthday parties. For the case of "getting a vehicle unstuck" false alarms are the videos involving cars and people in the car. However since the event of getting a vehicle unstuck is not fully performed they are counted as near-misses and not-relevant to that particular event. For the event of "grooming an animal" false alarms involve videos of animals such as a yellow python, a cat, a black and a white kitten, and a black dog. These animals are also involved with some sort of an interaction with people such as playing; however, not grooming. Our models cannot outperform the best trajectory features for some events such as "birthday party" (E006), and "parade" (E012). Given that our models focus on object based concepts, features exploiting temporal information work better in such events since they model the actions rather than the object dependencies. We plan to add temporal dependencies in our retrieval model; however, we have left this part for future work.

[20] fused all the low-level features; GIST, SIFT, cSIFT, mo-SIFT, STIP, DTF-HOG, and DTF-MBH that they have in their work and obtained very promising results. We also fused our results with DTF-HOG and DTF-MBH because they model the motion rather than the objects. In Table 2, we provide the fusion of low-level features provided in [20] as well as fusion of our models with HOG and MBH features. In almost all of the cases, fusion with our models outperforms the fusion of [20]. Only for events seven and eleven their fusion provides better scores. This is because MBH does not perform well for these events and hence the fusion results are worse. Perhaps this can be solved using a linear combination of features; however, we do not have their training sets so that we cannot create a validation set for tuning co-efficients. Instead we first standardize the scores of each individual feature set so that they have zero mean and unit variance. Then we take the arithmetic mean of the standardized scores and do the ranking using these scores.

For the events "birthday party," "flash mob gathering," "getting a vehicle unstuck," "grooming animal," "parade," "repairing an appliance," and "working on a sewing project", our models improve the results of low level features. For the event "parkour" the results are the same, and for the events "changing a vehicle tire," and "making a sandwich" fusion of low-level features is slightly better than fusion with our models. The results show that incorporating object-based concepts in addition to action concepts helps. Besides, higher-level object-based features are better than low-level object based features. In other words, our object-based concepts performs far better than the low-level object-based features such as SIFT, c-SIFT, and mo-SIFT. We suggest that action-based features should be combined with high-level object based features.

We see that fusion of HOG and MBH with our models improves the mean average scores. They[20] obtain an average of 24.4% MD rate at FA=5%; whereas, fusion with our models provide an average MD rate of 22.2% for the same FA. The lower MD rate, the more successful the model is.

In Figure 5 we provide sample video frames of the first five video retrievals for each event. In the figure there are five sample video frames on each row for an event. Relevance judgments for those videos are also provided. R means a relevant retrieval, and N means a video is not relevant to that event. Those might be near misses but the standard relevance judgment does not provide further information about such videos. The first five video retrievals are obtained using the fusion of HOG and MBH with our models. Note that even though some sample video frames seem to be relevant to an event, in the relevance judgment they might not be counted as relevant since somehow they might not satisfy the criteria of an event. For example, the second and fifth retrievals for the event *E006* seems to be examples of "birthday party"; however they are counted as not-relevant. We use the relevance judgment provided for evaluation of MED task so we did not judge the videos. When we look at the not-relevant retrievals in the figure we can see that they are mostly near-misses. Therefore, the next challenge is to differentiate relevant videos from near misses. One solution to dealing with this problem, after the initial ranking, can be creating another model to differentiate relevant videos from the near misses. Perhaps a tighter discriminative function should be employed since these videos are very close to each other in the feature space.

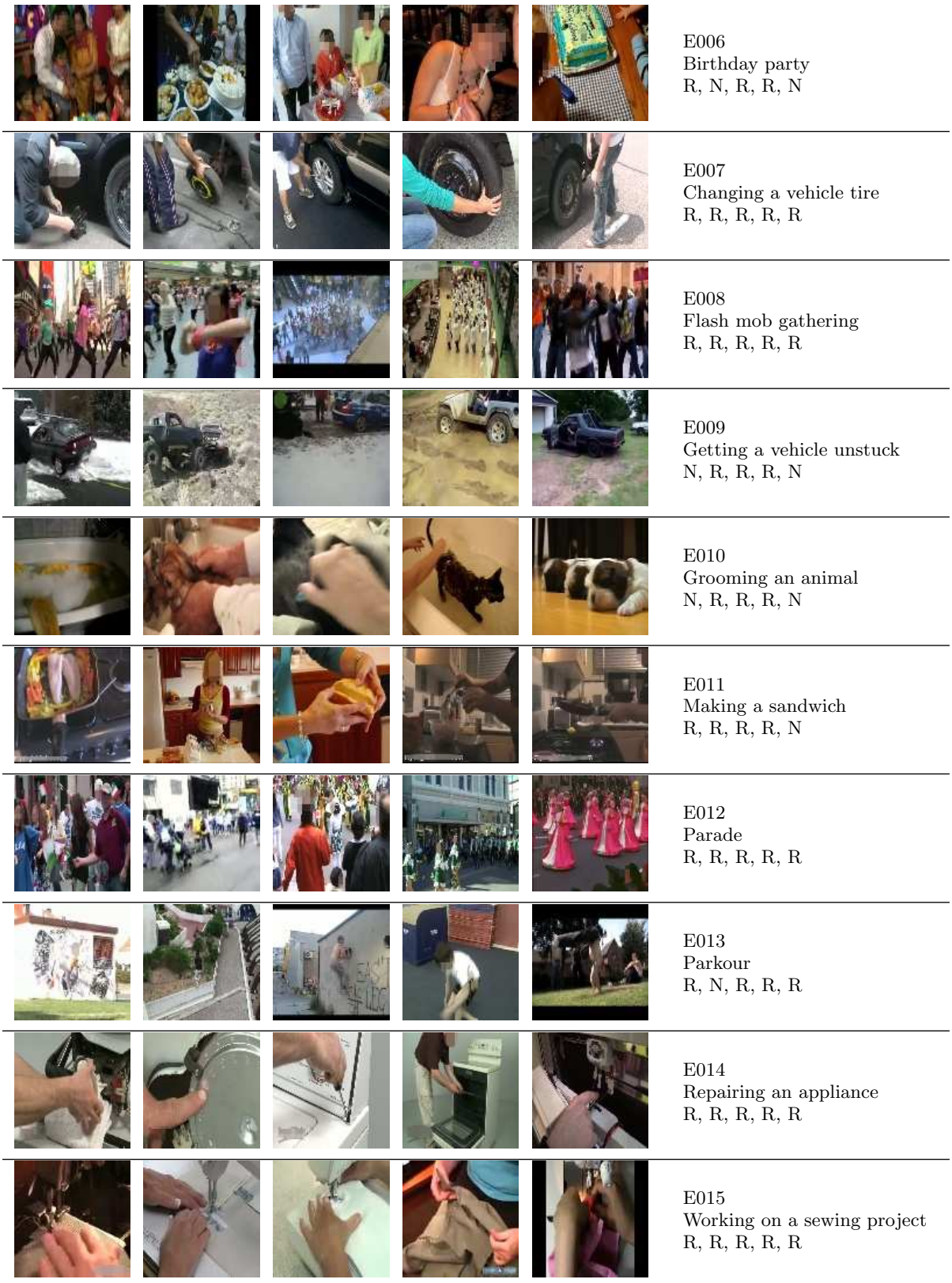

Figure 5: Sample video frames of the first five video retrievals for each event. Relevance judgment for each sample is respectively provided after the name of an event. R: Relevant, N: Not-relevant.

6. CONCLUSION

In this work, we model the dependencies of concepts for unconstrained videos in the context of multimedia information retrieval and particularly for event detection. We propose a method that builds MRF based models on top of concept detector outputs.

To this end, a number of concepts are created using the images. Then, we exploit the knowledge of the concepts in the video level to obtain a higher level of representation. We show that exploiting dependencies of a set of object-based concepts trained on static images can produce state-of-the-art results on event detection. We evaluate our models on a very large dataset consisting of approximately 45,000 unconstrained videos.

Even though the results of our model is promising, further improvements can be made to the model by incorporating other features for detecting concepts. The current model does not use any temporal information. Given that DTF-HoG and DTF-MBH are the best low-level temporal features, they can improve the performance of the model highly if incorporated within our models.

7. ACKNOWLEDGMENTS

This work was supported in part by the Center for Intelligent Information Retrieval and In part by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center contract number D11-PC20066. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. 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.

We would like to thank SRI-Sarnoff -Omar Javed, Jingen Liu, Hui Cheng, and Harpreet Sawhney- for providing us the rank lists for DTF-HOG, DTF-MBH, and STIP.

8. REFERENCES

[1] T. Althoff, H. O. Song, and T. Darrell. Detection bank: An object detection based video representation for multimedia event recognition. In *ACM Multimedia*, 2012.

[2] L. Ballan, M. Bertini, A. Bimbo, and G. Serra. Video event classification using bag of words and string kernels. In *ICIAP*, pages 170–178, 2009.

[3] I. ComScore. December 2013 U.S. online video rankings, 2014.

[4] S. Feng and R. Manmatha. A discrete retrieval model for image and video retrieval. In *CIVR*, pages 427–436, 2008.

[5] A. Habibian, K. van de Sande, and C. Snoek. Recommendations for video event recognition using concept vocabularies. In *ICMR*, 2013.

[6] ImageNet. An image database, 2013.

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

[8] Y. G. Jiang, X. Zeng, G. Ye, S. Bhattacharya, D. Ellis, M. Shah, and S. F. Chang. Columbia-ucf trecvid2010 multimedia event detection: Combining multiple modalities, contextual concepts, and temporal matching. In *NIST TRECVID Workshop*, 2010.

[9] T. Joachims. Optimizing search engines using clickthrough data. In *ACM KDD*, pages 133–142, 2002.

[10] L.-J. Li, H. Su, E. P. Xing, and L. Fei-Fei. Object bank: A high-level image representation for scene classification and semantic feature sparsification. In *NIPS*, 2010.

[11] J. Liu, Q. Yu, O. Javed, S. Ali, A. Tamrakar, A. Divakaran, H. Cheng, and H. Sawhney. Video event recognition using concept attributes. In *WACV*, pages 339–346, 2013.

[12] 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.

[13] M. Mazloom, E. Gavves, K. van de Sande, and C. Snoek. Searching informative concept banks for video event detection. In *ICMR*, pages 255–262, 2013.

[14] D. Metzler and W. B. Croft. A markov random field model for term dependencies. In *SIGIR*, pages 472–479, 2005.

[15] M. Naphade, J. Smith, J. Tesic, C. Shih-Fu, W. Hsu, L. Kennedy, A. Hauptmann, and J. Curtis. Large-scale concept ontology for multimedia. *Multimedia, IEEE*, 13(3):86 –91, 2006.

[16] 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*, pages 1298–1305, 2012.

[17] R. Poppe. A survey on vision-based human action recognition. *Image and vision computing*, 28(6):976–990, 2010.

[18] S. Sadanand and J. J. Corso. Action bank: A high-level representation of activity in video. In *CVPR*, pages 1234–1241, 2012.

[19] C. Snoek, M. Worring, J. van Gemert, J.-M. Geusebroek, and A. W. M. Smeulders. The challenge problem for automated detection of 101 semantic concepts in multimedia. In *ACM Multimedia*, pages 421–430, 2006.

[20] A. Tamrakar, S. Ali, Q., 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*, pages 3681–3688, 2012.

[21] D. Weinland, R. Ronfard, and E. Boyer. A survey of vision-based methods for action representation, segmentation and recognition. *Computer Vision and Image Understanding*, 115(2):224–241, 2011.

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

[23] Z. zhong Lan, L. Bao, S.-I. Yu, W. Liu, and A. G. Hauptmann. Double fusion for multimedia event detection. In *ACM Multimedia*, 2012.

[24] X. Zhou, X. Zhuang, S. Yan, S. Chang, M. Hasegawa-Johnson, and T. S. Huang. Sift-bag kernel for video event analysis. In *ACM Multimedia*, pages 229–238, 2008.