DISCOVER: Discovering Important Segments for Classification of Video Events and Recounting

Chen Sun and Ram Nevatia
University of Southern California, Institute for Robotics and Intelligent Systems
Los Angeles, CA 90089, USA
{chensun|nevatia}@usc.edu

Abstract

We propose a unified framework DISCOVER to simultaneously discover important segments, classify high-level events and generate recounting for large amounts of unconstrained web videos. The motivation is our observation that many video events are characterized by certain important segments. Our goal is to find the important segments and capture their information for event classification and recounting. We introduce an evidence localization model where evidence locations are modeled as latent variables. We impose constraints on global video appearance, local evidence appearance and the temporal structure of the evidence. The model is learned via a max-margin framework and allows efficient inference. Our method does not require annotating sources of evidence, and is jointly optimized for event classification and recounting. Experimental results are shown on the challenging TRECVID 2013 MEDTest dataset.

1. Introduction

We are in an era when shooting and sharing videos has never been easier. There are large amounts of unconstrained amateur videos which contain rich information but are poorly labeled. It is therefore important to build systems for video understanding and automatic video tagging. This paper focuses on the problem of high-level event classification and recounting for web videos. Given a query video, our framework provides not only a high level event label (*e.g. a wedding ceremony*), but also video segments which are important positive evidence and their textual descriptions (*e.g. people hugging*). The task is challenging for several reasons: Videos are shot by amateur users and are rather unstructured; the possibility of unobserved irrelevant video segments is high for query videos, adding noises for event classification. Due to the same reason, positive evidence can appear anywhere in the video, it is difficult to locate it by simple rules or a rigid temporal model. As an example, in a *flash mob* video, *dancing*, *marching* and *people cheering* can happen in different orders in an urban scene.

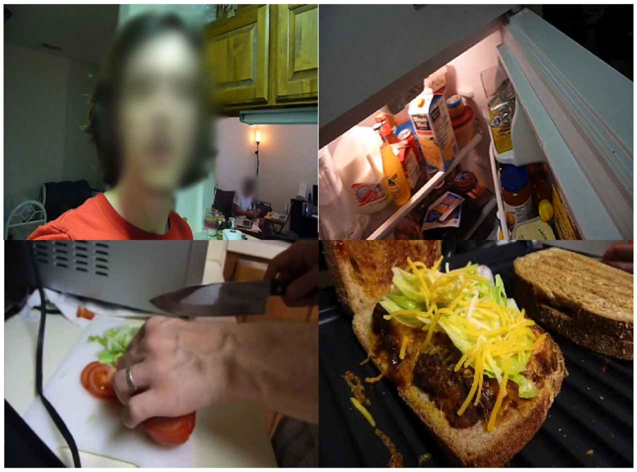


Figure 1. *(Top)* Randomly selected snapshots from a *making sandwich* video. *(Bottom)* Snapshots selected from the middle of each evidence clip identified by our framework from the same video.

Although there has been active research on event classification and recounting recently, the above challenges are not fully addressed. Low-level frameworks [9, 25] represent videos by spatio-temporally pooled feature vectors, requiring well aligned videos as the structure of spatio-temporal pyramids is rigid. [11, 20, 22] divide videos into clips uniformly to model the temporal structures within videos, they assume implicitly that videos are well cropped so that regions of interest are adjacent. [3] and [12] apply object and action detectors to generate video-level recounting. These methods do not explicitly identify regions of positive evidence for a specific event. This motivates us to discover important segments, classify high-level events and generate recounting jointly in a unified framework *DISCOVER*.

The underlying observation of DISCOVER is that the presence of high level event in videos is determined by the presence of positive evidence for that event. A positive evidence is characterized by a template of primitive actions and objects in a relatively short period of time. For example, humans can easily identify a *making a sandwich* video by the presence of *sandwich*, *pan* and *hand movement* within tens of frames, as shown in Figure 1.

The framework starts with primitive action based video representation with timestamps. It is obtained by segmenting videos into short clips and applying pre-trained primitive action classifiers to each clip. The presence of an event is determined by: (1) Global video representation, generated by pooling visual features over the entire video, (2) the presence of several pieces of different positive evidence which are consistent over time. We introduce an evidence localization model (ELM) which has a global template and a set of local evidence templates. ELM uses a compact action transition representation motivated by [20] to impose temporal consistency for adjacent pieces of evidence. The constraint is suitable for the usually diverse temporal structures of unconstrained web videos.

Inference in ELM is done by finding the best evidence locations which match the evidence templates well, and are consistent over time. It can be solved efficiently by dynamic programming. Once the pieces of evidence are located, we use the top weighted primitive actions as the descriptors for each evidence. ELM's parameters are learned via a max-margin framework described in [26]. We treat the locations of positive evidence as latent variables. The only supervised information required is video level event labels. We use the TRECVID 2013 MEDTest dataset [1] with 24,957 web videos for evaluation.

In summary, the main contributions of this paper are two fold: First, we propose a framework to jointly classify high level events, locate positive evidence and generate descriptions for unconstrained web videos; the method is efficient for large datasets during training and inference. Second, we show that our framework outperforms state-of-the-art methods on a challenging dataset, which confirms the validity of discovering positive evidence for event classification and recounting tasks.

2. Related Work

Most existing work on web video event classification addresses the problem without identification of important segments. In [15], the authors achieved good performance by careful selection of low level features and coding methods. It uses SIFT [13] and Dense Trajectories [25] low level visual features and Fisher Vector coding [16]. Video level features are aggregated over the entire videos. One obvious drawback is that each part of the video contributes equally to the final representation, making it prone to noise. Moreover, although spatio-temporal pyramids [9] can be built during feature aggregation, it requires the subcomponents of an event to happen in the same order, and be well aligned over time.

Several approaches try to introduce richer temporal models. [14] uses a tree structure with anchor positions to reward the presence of motion segments near the corresponding anchor positions. [22] uses variable-duration Hidden Markov Model to model a video event as a sequence of latent states with various durations. Video representations in the above approaches consist of low level features. [6] and [20] use *actom* or *action concept* as a mid-level representation and encode their temporal constraints. These approaches suffer from at least one of the following problems: (1) Temporal constraint is too rigid, (2) assumption that all segments are informative, (3) hard to locate positive evidence.

To find the discriminative parts of videos, [24] learns kernelized latent SVMs with no temporal constraints. [17] uses a simple algorithm to evaluate the quality of a possible cropping using other uncropped training data, the goal is to filter irrelevant segments from training dataset. [10] employs a dynamic pooling procedure by selecting informative regions for pooling low level features. These approaches ignore temporal structures which are important to event understanding.

For video description and event recounting, [7] uses captions as weak labels to learn an AND-OR graph based storyline. However, captions are usually not available for unconstrained web videos. [2] applies object tracks and body-postures to sentence generation; compared with our dataset, the dataset they use contains only a few objects and no camera motion. [3] assumes event classification is done and mines the co-occurrence of objects and bag of low level features.

3. Model

Our model consists of a **global template**, a set of local **evidence templates** and a **temporal transition constraint** of evidence set. Given a video, we find the sequence of video segments which achieve best overall score in matching the evidence templates and meeting the temporal constraints. An event label is assigned based on the global feature of a video, as well as features from the selected pieces of evidence. An illustration is given in Figure 2.

Our model is related to the Deformable Part-based Model (DPM) [5] in the sense that they both try to find discriminative components from query data. However, our model is motivated by locating pieces of positive evidence for event classification and recounting, thus the representation and constraints are different.

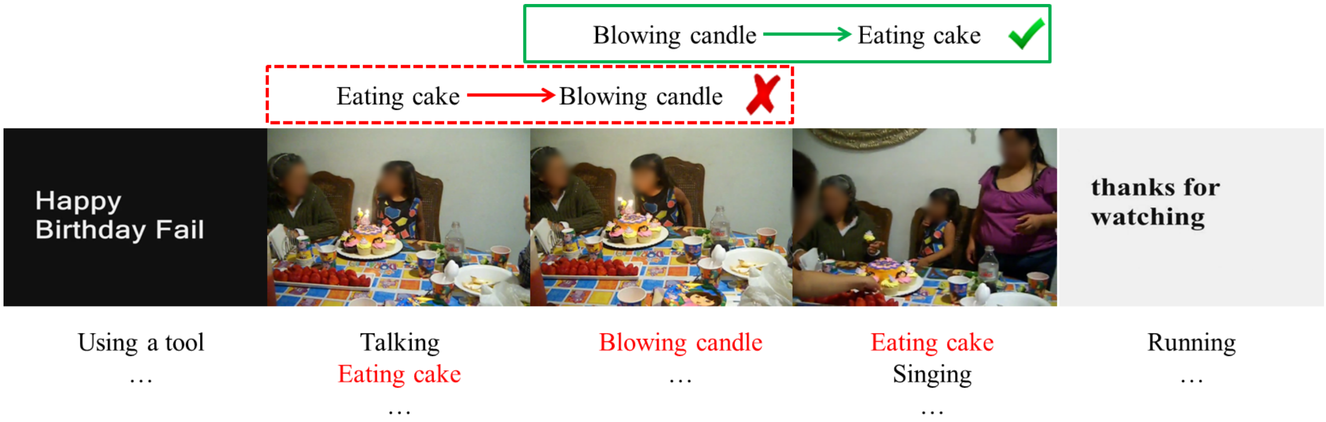


Figure 2. (*Middle*) A typical *birthday party* video from the dataset, some of the video clips are irrelevant to the event. (*Bottom*) Each clip has a vector of primitive action classifier responses, where the highest scoring ones are listed. Primitive actions in red have high weights in evidence templates. (*Top*) Two configurations of evidence locations. The green one scores higher than the red one, as transition from eating cake to blowing candle is highly unlikely.

3.1. Video Representation

Videos are represented by a set of primitive action responses with timestamps.

Given an input video V, we first divide it into a sequence of short clips $[S_1\ S_2\ ...\ S_n]$. This can be done by a sliding window with uniform size, or by shot boundary detection. We then apply a set of pre-trained action classifiers, where each action classifier maps a video clip to a confidence score on how likely the action appears in the clip, given by

$$c_{ji} = f_i(S_j) \tag{1}$$

This representation can be easily extended to objects, which are image or patch based, by pooling classifier outputs sampled from S_j.

Once this step is finished, video V is represented by a matrix $\mathcal{C} = [c_{j,i}]$. The j-th row of the matrix gives a complete action based description for S_j.

It is reasonable to expect that to generate meaningful video descriptions, certain action types are necessary. Meanwhile, action classifiers can also be seen as nonlinear projections of original feature space, which provide discriminative information for classification [23, 20]. In this paper, we use action annotations from training videos as well as independent dataset. Some of the actions are not related to high level events. A wide range of object and scene types can also be used, but not addressed in this paper.

3.2. Evidence Localization Model

ELM's evidence templates are learned and used for evidence localization from query videos. For example, *vehicle moving*, *people dancing* and *people marching* are all related to the *parade* event and may be expected to appear in different videos. Meanwhile, if a video contains *people dancing* and *people marching*, it is highly likely to be a *parade* event. These two primitive actions should have high weights in their corresponding evidence templates. However, in practice the locations of segments providing evidence are usually not available for training. We address this problem by treating the locations as latent variables, and define the scoring function

$$f_{\mathbf{w}}(\mathcal{C}) = \max_{\mathbf{z} \in \mathcal{Z}} [f_r(\mathcal{C}) + f_p(\mathcal{C}, \mathbf{z}) + f_t(\mathcal{C}, \mathbf{z})] \tag{2}$$

where $\mathcal{C}$ is action response matrix of a video, each row of $\mathcal{C}$ corresponds to a video segment, $\mathcal{Z}$ is the set of all possible configurations of evidence locations. $f(\mathcal{C})$ can be decomposed into the following terms

Global score $f_r(\mathcal{C})$ measures event similarity based on global video features. This can be done by extracting statistics information from all clips of a video. For example, one can use average pooling technique

$$h_r(\mathcal{C}) = \frac{1}{N} \sum_{i=1}^{N} \mathcal{C}_i$$

where N is the number of rows in $\mathcal{C}$ and $\mathcal{C}_i$ is the i-th row of $\mathcal{C}$.

Global term can then be expanded to

$$f_r(\mathcal{C}) = \mathbf{w_r}^\intercal h_r(\mathcal{C}) \tag{3}$$

parameterized by global template $\mathbf{w_r}$.

Local evidence score $f_p(\mathcal{C}, \mathbf{z})$ measures how the located pieces of evidence matches evidence templates. Denote z_i as the clip index of i-th evidence template, and T as the total number of evidence templates in evidence set, we have

$$f_p(\mathcal{C}, \mathbf{z}) = \sum_{i=1}^{T} \mathbf{w_{p,i}}^\intercal \mathcal{C}_{z_i} \tag{4}$$

The vector $\mathbf{w_{p,i}}$ can be seen as an evidence template containing the desired action responses. If action classifiers are perfect, $\mathbf{w_{p,i}}$ should be sparse as only a small subset of actions should appear in an evidence. However, since the current state-of-the-art action classifiers are far from perfect, we decide not to impose sparsity constraint on $\mathbf{w_p}$.

Temporal consistency score $f_t(\mathcal{C}, \mathbf{z})$ evaluates the validity of the selected pieces of evidence over time. Considering an ideal scenario for *getting vehicle unstuck* event, it is obvious that *pushing vehicle* should happen before *vehicle moving*.

In this paper, we use the Hidden Markov Model Fisher Vector (HMMFV) [20] to model temporal consistency. The basic idea is to encode the action transitions with the Fisher kernel technique. Given a trained HMM parameterized by Θ, and a collection of action response data $\mathcal{C}$ from clips of a video, each dimension $\varphi_{i,j}(\mathcal{C})$ corresponds to the partial derivative of the log-likelihood function $\log P(\mathcal{C}|\Theta)$ over transition parameter $\theta_{i,j}$, given by

$$\varphi_{i,j}(\mathcal{C}) = \sum_{t=1}^{N} \alpha_{t-1}(j)\left[c_{t,i}\beta_t(i) - \beta_{t-1}(j)\right] \tag{5}$$

where N is the number of clips, $c_{t,i}$ is the emission probability of action i at clip t, i and j are action types. $\alpha_t(i)$ is the probability of observing first t clips and the t-th clip belongs to i-th primitive action, and $\beta_t(i)$ is the probability of observing clips from the $(t+1)$-th to the end given the t-th clip belongs to the i-th primitive action. αs and βs can be computed efficiently via dynamic programming.

Assuming the positions of evidence $\mathbf{z}$ are sorted in temporal order $[z_{t_1} \ldots z_{t_T}]$, and a uniform prior distribution for all actions, we have

$$f_t(\mathcal{C}, \mathbf{z}) = \sum_{i=1}^{T-1} \mathbf{w_t}^{\intercal} \varphi([\mathcal{C}_{z_{t_i}}; \mathcal{C}_{z_{t_{i+1}}}]) \tag{6}$$

which measures the compatibility of all adjacent pairs of evidence.

3.3. Compact Temporal Constraint

One potential problem of the above temporal constraint is that the dimension of φ grows quadratically with the number of actions, which makes it computationally infeasible to support a large vocabulary. We use the following guideline to select a subset of actions: considering evidence templates $\mathbf{w_{p,i}}$ $(i = 1, \ldots, T)$, higher values in $\mathbf{w_{p,i}}$ reflect the dimensions that are important components for the evidence template. We therefore select a subset of actions by taking the average of $\mathbf{w_{p,i}}$ $(i = 1, \ldots, T)$, and picking the actions corresponding to D dimensions with highest values.

Our temporal constraint is flexible and data-driven: by learning the parameter vector $\mathbf{w_t}$ from training data, it can be used for both evidence sets with more rigid structures as well as those with no clear temporal orders.

4. Inference

Inference involves solving Equation 2 by finding the assignment of latent variables $\mathbf{z}$ which maximizes $f_{\mathbf{w}}(\mathcal{C})$. We rewrite Equation 2 into

$$f_{\mathbf{w}}(\mathcal{C}) = \psi_1(\mathcal{C}) + \max_{\mathbf{z} \in \mathcal{Z}} g_{\mathbf{w}}(\mathcal{C}, \mathbf{z}) \tag{7}$$

where

$$g_{\mathbf{w}}(\mathcal{C}, \mathbf{z}) = \sum_{i=1}^{T} \psi_2(\mathcal{C}, z_i, t_i) + \sum_{i=1}^{T-1} \psi_3(\mathcal{C}, z_i, z_{i+1})$$

Here T is the number of evidence templates, z_i is the location of i-th evidence in temporal order, and t_i is the evidence template index.

The problem now becomes one of selecting T pieces of evidence sequentially from a set of N video clips where the choice of the i-th evidence is only affected by the $(i-1)$-th evidence. Let $G(i, z, t)$ be the maximum score by selecting the first i evidence locations given that the i-th location is z and its template index is t. We have

$$\begin{aligned} G(i, z, t) = &\max_{z', t'} [G(i-1, z', t') + \psi_3(\mathcal{C}, z', z)] \\ &+ \psi_2(\mathcal{C}, z, t) \end{aligned}$$

The target score is $\max_{z,t} G(T, z, t)$, which can be solved in $O(T^3 N^2)$ by dynamic programming. Positive evidence locations can be obtained by backtracking.

In practice, we find retaining top $M = 10$ candidate evidence locations already provides good performance. This reduces the time complexity to $O(TM^2)$.

5. Learning

Our model is parameterized by global template $\mathbf{w_r}$, evidence templates $\mathbf{w_p}$, temporal constraint vector $\mathbf{w_t}$ and transition parameters of HMM. We first introduce the learning of the first three vectors $\mathbf{w} = [\mathbf{w_r}\ \mathbf{w_p}\ \mathbf{w_t}]$, and describe the last one in Section 5.1.

Given labeled training set $\{\mathcal{C}_i, y_i\}$, $y_i \in \{-1, 1\}$, $i = 1, 2, \ldots, N$, we learn the parameters $\mathbf{w}$ by max-margin criteria similar to [5] and [26]

$$\min_{\mathbf{w}, \xi} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{N} \xi_i \tag{8}$$

$$\text{s.t. } y_i f_{\mathbf{w}}(\mathcal{C}) \geq 1 - \xi_i, \xi_i \geq 0, \forall i$$

ξ is a vector of slack variables, it measures the degree of misclassification. C is a cost variable which balances the two terms in target function.

The optimization problem is semi-convex. We use the quadratic programming based solver proposed in [26].

5.1. Initialization

Since Equation 8 poses a semi-convex problem, initialization plays an important factor in getting a good local optimum solution. We determine the number of evidence templates and the evidence locations in positive training videos, in the following steps:

1. Select all action response vectors from positive training data to form a set D_p, and randomly select action response vectors from negative training data to form a set D_n.

2. Do a K-Means clustering on D_p with a large K (*e.g.* 200).

3. For each cluster, use entries of D_p from that cluster as positive set and D_n as negative set to train a linear SVM classifier. Apply the classifier to all action response vectors from training set. A video level score is produced by taking the maximum score of its action response vectors. Compute the average precision.

4. Pick a cluster with high average precision as an evidence template, given certain criteria are met. Stop if the maximum evidence template size is reached.

5. For each selected cluster, apply its classifier to each positive video, select the clip with the highest response as the evidence location.

We use two criteria for picking evidence templates: minimum average precision (minAP) and maximum percentage of evidence overlap (maxEO). A cluster is picked if the average precision of its corresponding classifier is higher than minAP, and the overlapping percentage of its selected evidence locations with previous selected locations is lower than maxEO.

The intuition behind this initialization method is that if a classifier trained by positive action responses from a cluster performs well over the whole dataset, then the corresponding clips are likely to be representative for most of the positive videos.

After initialization, we use the evidence templates to select the top actions used for modeling temporal constraint. A general HMM for selected actions is learned by sampling training data.

6. Event Recounting

Generating video description for recounting is straightforward in our framework. After running the inference algorithm, we have all the evidence locations in a video. A summary of the video can then be generated by ordering the evidence temporally. To generate textual descriptions for a evidence clip, we use actions with top responses weighted by evidence template. Suppose $\mathbf{w_{p,i}} = [w_{i,1} \ldots w_{i,D}]$ is the i-th evidence template, and $\mathcal{C}_j = [c_{j,1} \ldots c_{j,D}]$ is the action response vector. We select the top action given by

$$\arg\max_k w_{i,k} c_{j,k} \tag{9}$$

Another possible strategy for video description is to learn linear event classifiers from average pooled action responses. The classifiers can be used as pseudo evidence templates, one for each event [19]. Compared with ELM, this global strategy lacks the diversity of having multiple evidence types per event. We compare these two strategies in Section 7.3.

7. Experiments

This section describes the dataset we used for evaluation, as well as evaluation results for classification and recounting tasks.

7.1. Dataset

We used TRECVID 2013 Multimedia Event Detection dataset [1] for evaluation. The dataset contains unconstrained web videos varying in length, quality and resolution. We chose to evaluate the ten events listed in Table 1.

We used three different partitions: *Background*, which contains 4,992 background videos not belonging to any of the target events; *100EX*, which contains 100 positive videos for each event; *MEDTest*, which contains 24,957 videos. To train a model for a specific event, we used all background videos from *Background*, and positive videos from *100EX* of that event only. Videos in *MEDTest* were used for testing.

We used two datasets UCF 101 [18] and MED action annotation set [8] to learn primitive action classifiers. UCF 101 has 13,320 videos from 101 categories, the videos are of similar quality to TRECVID 2013 MED dataset, but many of the action types are not related to the 10 events. MED action annotation set has 60 action types annotated directly on *100 EX* videos, the actions are highly related to events.

7.2. Classification Task

Classification task is to assign a single event label to each query video. We report our results in average precision (AP).

To learn primitive action classifiers, we used Dense Trajectories (DT) features [25], and obtained video level representations by Fisher Vector coding [21]. LIBLINEAR[4] was used for SVM classifier training. Action classifiers were applied to video for every 100 frames, with a 50 frame step size. We used a two-fold cross validation to select our framework's parameters, which includes the cost C and relative weight of positive and negative samples. For model initialization, we set the size of candidate clusters to 200, minAP to 0.1, and maxEO to 20%. We selected top 40 primitive actions for temporal constraints.

Comparison with Baselines. As the overall performance is affected by the choice of the global video representation, we first trained our evidence localization model

Event name	ID	Global	ELM_NT	ELM
Birthday party	6	13.6	**17.4**	17.1
Changing a vehicle tire	7	8.7	13.1	**17.7**
Flash mob gathering	8	31.2	42.7	**57.3**
Getting a vehicle unstuck	9	21.9	**28.3**	25.2
Grooming an animal	10	7.9	11.4	**14.9**
Making a sandwich	11	5.1	11.4	**13.2**
Parade	12	26.2	**33.8**	**33.7**
Parkour	13	19.4	39.9	**43.6**
Repairing an appliance	14	4.2	17.0	**20.6**
Sewing project	15	6.8	17.0	**24.2**
mean Average Precision (mAP)		14.5	23.2	**26.8**

Table 1. Average precision comparison among global baseline, ELM without temporal constraint and the full ELM on *MEDTest*

ID	HMMFV	ELM+HMMFV	DTFV	ELM+DTFV
6	**24.2**	22.7	19.4	**22.0**
7	14.7	**19.4**	17.1	**25.2**
8	52.9	**59.7**	55.7	**61.5**
9	29.6	**34.0**	35.6	**38.1**
10	8.9	**11.4**	12.7	**15.3**
11	17.1	**18.2**	15.4	**17.1**
12	32.6	**37.3**	33.3	**37.8**
13	53.5	**54.9**	55.4	**57.1**
14	25.7	**28.1**	**37.1**	36.2
15	15.0	**25.0**	19.1	**27.9**
mAP	27.4	**31.1**	30.1	**33.8**

Table 2. Performance gain by incorporating two state-of-the-art global video representations into our framework

(ELM) *without* using the global term. We chose global average pooling of the action responses (**Global**) as the first baseline, defined by

$$h(\mathbf{X}) = \frac{1}{N}\sum_{i=1}^{N} X_i$$

where $\mathbf{X}$ is a N by D matrix, N is the number of total action response vectors, D is the number of actions. X_i is the i-th row of $\mathbf{X}$. We used linear kernel SVM, and 5-fold cross validation to select classifier parameters.

To validate the effectiveness of temporal constraint term, we also provided results of ELM with no temporal constraint (**ELM_NT**). The results are shown in Table 1.

From the table, it is easy to see that ELM has much higher mean average precision compared with the global baseline. One possible explanation is that, by locating positive evidence instead of treating all clips as being equally important, ELM is more robust against various irrelevant segments in testing videos. The results also indicate that the temporal constraint is effective for most of the events.

Working with State-of-the-Art Global Methods. Any global event classification approach that provides a fixed length video level feature vector can be incorporated into our framework as the global term. It is interesting to see if our evidence based framework can improve these methods. Recently, [15] showed that using multiple types of low level features can increase event classification performance. It is then important to fix the low level features to compare different frameworks.

We chose two different state-of-the-art approaches following this rule. The first one uses an action transition based representation called *HMMFV* [20]. It accumulates action transition statistics over the entire video. We used the same set of pre-trained action classifiers as those used in our approach. We refer to this method as **HMMFV**.

We also implemented a low-level based framework based on Dense Trajectories. The features were extracted at step size of 10. We used Fisher Vector with both first and second order terms, and followed the suggestions in [21] to apply power normalization and $l2$-normalization to the feature vectors. We used PCA to reduce the dimension of DT features to 128, and used a codebook size of 64. We refer to this method as **DTFV**.

Both methods used a linear SVM classifier, the parameters were selected by 5-fold cross validation. The two global approaches as well as our ELM method were all based on DT features alone.

According to Table 2, by discovering evidence from videos, ELM achieves significant and consistent improvement over HMMFV and DTFV.

7.3. Recounting Task

In TRECVID Multimedia Event Recounting (MER) task, a video description is defined as a video snippet with a starting frame, an ending frame and a textual description. We used the ELM framework to obtain such snippets, and compared it with the global strategy described in Section 6. The number of snippets selected per video was fixed to ELM's evidence set size $T = 2$ for both strategies.

The evaluation of video recounting results is difficult, as there is no groundtruth information for which snippets are correct; to the best of our knowledge, there is also little previous work to compare with. We conducted an experiment based on human evaluation. Eight volunteers were asked to serve as evaluators. Before evaluation, each evaluator was shown the event category descriptions in text, as well as one or two positive examples in the training set. For each event, 10 positive videos were picked randomly from *MEDTest*. However, we informed evaluators that negative videos may also appear to avoid biased prior information. Evaluators were first presented with snippets generated by ELM to assign event labels, and then presented with snippets generated by the global strategy from the same video to compare which snippets are more informative and whose descriptions of the snippets are more accurate. Two criteria were

Average video length	172.9 seconds
Average snippet length	7.1 seconds
Ratio	**4.1%**
Average Accuracy	**86.3%**

Table 3. The ratio of our method's average snippet length over average video length, and the average accuracy of labels from evaluators

Event	Better	Similar	Worse
Birthday party	**4**	4	2
Changing a vehicle tire	**7**	0	3
Flash mob gathering	**5**	1	4
Getting a vehicle unstuck	3	**5**	2
Grooming an animal	**8**	0	2
Making an sandwich	**6**	0	4
Parade	**5**	2	3
Parkour	4	0	**6**
Repairing an appliance	**8**	0	2
Sewing project	3	**5**	2
Total	**53**	33	14

Table 4. Evaluators' comparison of ELM over global strategy. Assignments of *better*, *similar* and *worse* were aggregated via average

used: average accuracy, which measures the percentage of correctly labeled snippets; and relative performance, counting evaluators' preference of video-level recounting results generated by the two approaches.

Table 3 shows the average length of videos and snippets, as well as average accuracy. ELM achieves 86.3% average accuracy by selecting only 4% of frames in the original videos. This shows that our approach provides reasonable good snippets for users to rapidly and accurately grasp the basic idea of video events.

Table 4 summarizes the evaluators' preferences between ELM and the global strategy for each event. It can be seen that ELM is better for most of the events. Several recounting results are shown in Figure 3, where snippets generated by ELM are on the left. Among the three examples, our *flash mob gathering* snippets provide more diverse but also related information. Global strategy failed to assign proper description to the *repairing an appliance* video, where hands are not present in the selected snippets. The bottom row is an example of *making a sandwich* video where ELM's output is worse.

Most of the selected actions for description came from the MED action annotation set. This indicates the benefit of using event related actions to build the vocabulary.

8. Conclusion

This paper proposes the DISCOVER framework for video event classification and recounting. It classifies unconstrained web videos by discovering important segments characterized by primitive actions and their transitions. DISCOVER allows efficient learning and inference, and is generalizable to using objects and scenes. Experimental results show that it outperforms current state-of-the-art classification methods on a challenging large scale dataset. For event recounting, DISCOVER locates important segments which is seldom addressed in previous work. It also has the potential of event detection, it is a topic to be explored in future work.

9. Acknowledgement

This work was supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center contract number D11PC0067. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes nonwithstanding 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. Computation for the work described in this paper was supported by the University of Southern California Center for High-Performance Computing and Communications (hpcc.usc.edu).

References

[1] http://www.nist.gov/itl/iad/mig/med13.cfm. 2, 5

[2] A. Barbu et al. Video in sentences out. In *UAI*, 2012. 2

[3] P. Das, C. Xu, R. F. Doell, and J. J. Corso. A thousand frames in just a few words: Lingual description of videos through latent topics and sparse object stitching. In *CVPR*, 2013. 1, 2

[4] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. LIBLINEAR: A library for large linear classification. *JMLR*, 2008. 5

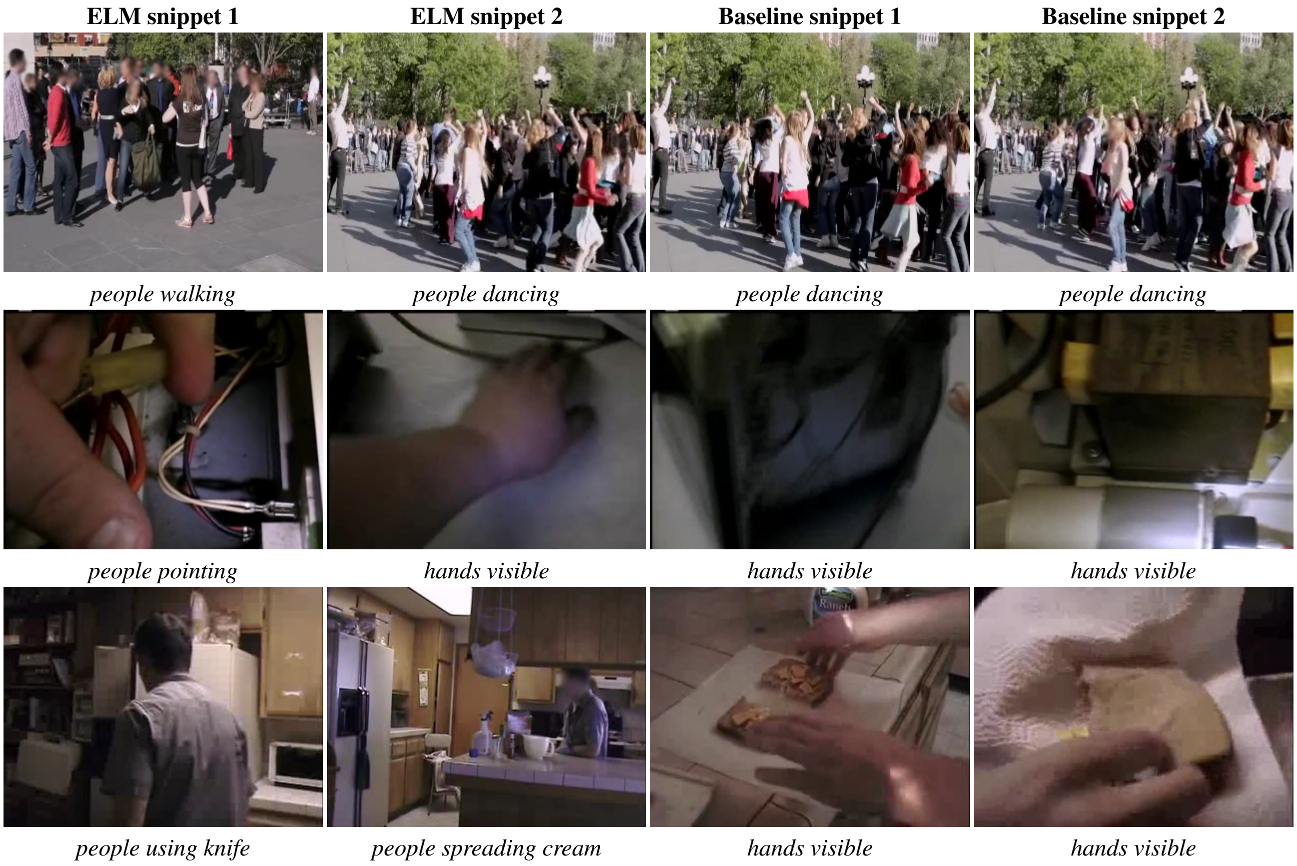
[5] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. In *PAMI*, 2009. 2, 4

[6] A. Gaidon, Z. Harchaoui, and C. Schmid. Actom sequence models for efficient action detection. In *CVPR*, 2011. 2

[7] A. Gupta, P. Srinivasan, J. Shi, and L. S. Davis. Understanding videos, constructing plots learning a visually grounded storyline model from annotated videos. In *CVPR*, 2009. 2

[8] H. Izadinia and M. Shah. Recognizing complex events using large margin joint low-level event model. In *ECCV*, 2012. 5

[9] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *CVPR*, 2008. 1, 2


ELM snippet 1

ELM snippet 2

Baseline snippet 1

Baseline snippet 2

people walking

people dancing

people dancing

people dancing

people pointing

hands visible

hands visible

hands visible

people using knife

people spreading cream

hands visible

hands visible

Figure 3. Event recounting results generated by ELM and the baseline approach. The video events are *flash mob gathering*, *repairing an appliance* and *making a sandwich* respectively. ELM was labeled as *better* than baseline by evaluators for the top two recounting results.

[10] W. Li, Q. Yu, A. Divakaran, and N. Vasconcelos. Dynamic pooling for complex event recognition. In *ICCV*, 2013. 2

[11] W. Li, Q. Yu, H. Sawhney, and N. Vasconcelos. Recognizing activities via bag of words for attribute dynamics. In *CVPR*, 2013. 1

[12] J. Liu, Q. Yu, O. Javed, S. Ali, A. Tamrakar, A. Divakaran, H. Cheng, and H. S. Sawhney. Video event recognition using concept attributes. In *WACV*, 2013. 1

[13] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 2004. 2

[14] J. C. Niebles, C.-W. Chen, , and L. Fei-Fei. Modeling temporal structure of decomposable motion segments for activity classification. In *ECCV*, 2010. 2

[15] D. Oneata, J. Verbeek, and C. Schmid. Action and Event Recognition with Fisher Vectors on a Compact Feature Set. In *ICCV*, 2013. 2, 6

[16] F. Perronnin and C. Dance. Fisher kernels on visual vocabularies for image categorization. In *CVPR*, 2007. 2

[17] S. Satkin and M. Hebert. Modeling the temporal extent of actions. In *ECCV*, 2010. 2

[18] K. Soomro, A. R. Zamir, and M. Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. *CRCV-TR-12-01*. 5

[19] C. Sun, B. Burns, R. Nevatia, C. Snoek, B. Bolles, G. Myers, W. Wang, and E. Yeh. ISOMER: Informative segment observations for multimedia event recounting. In *ICMR*, 2014. 5

[20] C. Sun and R. Nevatia. ACTIVE: Activity concept transitions in video event classification. In *ICCV*, 2013. 1, 2, 3, 4, 6

[21] C. Sun and R. Nevatia. Large-scale web video event classification by use of fisher vectors. In *WACV*, 2013. 5, 6

[22] K. Tang, F.-F. Li, and D. Koller. Learning latent temporal structure for complex event detection. In *CVPR*, 2012. 1, 2

[23] L. Torresani, M. Szummer, and A. W. Fitzgibbon. Efficient object category recognition using classemes. In *ECCV*, 2010. 3

[24] A. Vahdat, K. Cannons, G. Mori, I. Kim, and S. Oh. Compositional models for video event detection: A multiple kernel learning latent variable approach. In *ICCV*, 2013. 2

[25] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu. Dense trajectories and motion boundary descriptors for action recognition. *IJCV*, 2013. 1, 2, 5

[26] Y. Yang and D. Ramanan. Articulated pose estimation with flexible mixtures-of-parts. In *CVPR*, 2011. 2, 4