Minimally Needed Evidence for Complex Event Recognition in Unconstrained Videos

Subhabrata Bhattacharya Felix X. Yu Shih-Fu Chang
Digital Video and Multimedia Lab
Columbia University, New York, USA
{subh,yuxinnan,sfchang}@ee.columbia.edu

ABSTRACT

This paper addresses the fundamental question – How do humans recognize complex events in videos? Normally, humans view videos in a sequential manner. We hypothesize that humans can make high-level inference such as an event is present or not in a video, by looking at a very small number of frames not necessarily in a linear order. We attempt to verify this cognitive capability of humans and to discover the Minimally Needed Evidence (MNE) for each event. To this end, we introduce an online game based event quiz facilitating selection of minimal evidence required by humans to judge the presence or absence of a complex event in an open source video. Each video is divided into a set of temporally coherent microshots (1.5 secs in length) which are revealed only on player request. The player's task is to identify the positive and negative occurrences of the given target event with minimal number of requests to reveal evidence. Incentives are given to players for correct identification with the minimal number of requests.

Our extensive human study using the game quiz validates our hypothesis - 55% of videos need only one microshot for correct human judgment and events of varying complexity require different amounts of evidence for human judgment. In addition, the proposed notion of MNE enables us to select discriminative features, drastically improving speed and accuracy of a video retrieval system.

Category and Subject Descriptors: H.4 [Information Systems Applications] : Miscellaneous

Keywords: Minimally Needed Evidence, Complex Event recognition, Multimedia Event Detection, Video Retrieval

1. INTRODUCTION

Recognition of complex events in consumer videos such as "parade" or "changing a tire" is propelling a new breed of research [2, 3, 11, 12, 17, 18, 23, 28, 31] in multimedia and computer vision. This is an extremely challenging problem [2,3,8] as it involves high level machine understanding of videos that are semantically diverse, and prone to frequent illumination changes, large background clutter, and significant camera motion.

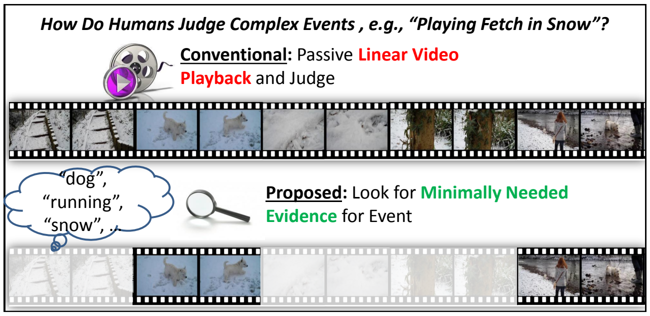

How Do Humans Judge Complex Events , e.g., "Playing Fetch in Snow"?
Conventional: Passive Linear Video Playback and Judge
"dog", "running", "snow", ..
Proposed: Look for Minimally Needed Evidence for Event

Figure 1: Minimally Needed Evidence (MNE) for identifying a complex event in a video.

Although impressive research has been conducted in this direction, they usually are based on a common paradigm that employs sequential processing of information in videos. Quite interestingly, human perception of events from videos however, may be achieved with a completely different mechanism. An analogous problem is recognition of objects in images. Recent studies [7, 10] suggest human recognition operates in a non-linear mechanism resulting in an excellent performance.

This motivates us to investigate a fundamental question – How do humans recognize complex events and how much evidence is needed for successful recognition? By studying the event recognition process employed by humans, we can explain vital cognitive aspects of human understanding of high-level events, and consequently find interesting differences between humans and machines in the same process. This can potentially help improve machine-based event recognition performance.

To this end, we develop the *Event Quiz Interface (EQI)*, an interactive tool mimicking some of the key strategies employed by the human cognition system for high-level understanding of complex events in unconstrained web videos. EQI enables us conduct an extensive study on human perception of events through which we put forth the notion of *Minimally Needed Evidence (MNE)* to predict the presence or absence of an event in a video. MNEs as shown in Fig. 1, are a collection of *microshots* (set of *spatially* downsampled contiguous frames) from a video. Note that our emphasis on scaled-down version of original frames, is coherent to the term *Minimal*, in this context. Our study reinforces our hypothesis that in a majority of cases humans make correct decisions about an event with just a small number of frames. It reveals that a single microshot serves as an MNE in 55% of test videos to identify an event correctly and 68% of videos can be correctly rejected after humans had viewed only a single microshot.

We make the following technical contributions in this paper: (a) We propose a novel game based interface to study human cogni-

WOODSTOCK '14, Glasgow, UK

tion of video events (Section 3.1), which can also be used in a variety of tasks in complex event recognition that involve human supervision [3, 17, 23, 28, 31], (b) Based on our extensive study, we demonstrate that humans can recognize complex events by seeing one or two chunks of spatially downsampled shots, less than a couple seconds in length (Section 3.2), (c) We leverage on positive and negative visual cues selected by humans for efficient retrieval (Section 4), and finally, (d) We perform conclusive experiments to demonstrate significant improvement in event retrieval performance on two challenging datasets released under TRECVID (Section 5) using off-the-shelf feature extraction techniques applied on MNEs versus evidence collected sequentially.

In addition, we further conjecture that the utility of MNEs can be perceived beyond complex event recognition. They can be used to drive tasks including feature extraction, fine-grained annotation and detection of concepts. Furthermore, they can also be used to investigate if underlying temporal structure in a video is useful for event recognition.

2. RELATED WORK

Video event recognition is a widely studied research area in the field [8] of multimedia and computer vision. That said, some of the noted work in recent past, pertinent to recognition in *unconstrained settings* include machine interpretation of either low-level features [13, 26] directly extracted from human labeled event videos [18, 20] or training intermediate-level semantic concepts that require expensive human annotation [1,17] or a combination of both [9,21].

All these automated efforts require a temporal scan of video content to initiate processing. In practice, videos are first divided into small temporal segments uniformly or using some scene change detection techniques before, further processing. Some sophistication on selection of relevant shots can be achieved through pooling [2,11], using latent variables [23] or modeling attribute dynamics [12] that preserve the temporal structure present in the videos.

Techniques involving full human interaction [22, 27, 30] to perform video retrieval, have also been studied. Efforts in this direction derive a strong similarity with human annotation of videos [5, 25, 29], as both tasks necessitate linear playback of a video by humans. While this paradigm may be suitable in the persistent video surveillance perspective, we argue that for recognition of complex events, it is an over utilization of resources.

We propose a completely different perspective complex recognition tasks. Specifically, our approach aims to identify minimal evidence required for recognition with or without a temporal order. In this context, our work draws some level of conceptual similarities with [6] and [4], albeit both are in different domains. In [6], the authors propose a max-margin framework to identify the number of frames needed to detect simple events e.g. "smile". Likewise, the authors of [4] introduce a game based interface to collect discriminative regions and features corresponding to them that humans see in an image to identify an object.

Our work also aligns with efforts that apply computational human attention models [7, 10, 14] or visual saliency [15] to solve high-level visual recognition or video summarization tasks. However, these methods, do not aim to find the minimal evidence and they are not designed for recognition of complex events. Needless to say, they are also computationally prohibitive.

3. EVIDENCE BASED RECOGNITION

In order to facilitate video playback in a non-linear fashion, we resort to a technique illustrated in Fig. 2. The objective of this interface is to enable humans discover Minimally Needed Evidences for the recognition of complex events. A more detailed implementation is provided in the later sections of the paper. Our interface design conforms to *minimalist principles* so that whatever human efforts involved can be judiciously utilized. Instead of asking humans to directly select the most discriminative shots from the entire video of the event in question, we use only a small set of *interesting microshots*. Compared to keyframes, microshots can reveal local temporal structure in videos, and therefore provide important cues about major human actions, in addition to contextual evidence such as scenes (outdoor/indoor) etc. Also, in contrast to regular shots, these are more compact, do not require sophisticated boundary detection techniques which can add to the computational overhead. More details on the microshot extraction is provided in the implementation section (Section 5.2).

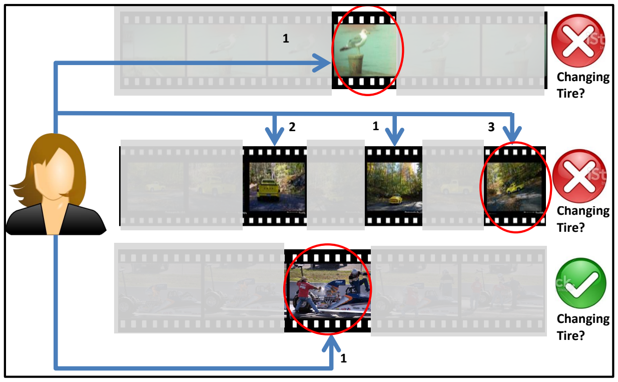


Figure 2: Collecting Minimally Needed Evidences (MNEs) to identify complex events in videos: A player watches a microshot (revealed only after request) to decide whether an event is present or absent in a video. Numbers over each microshot indicate the order in which they are revealed with microshots encircled in red are revealed just before the player made a decision.

3.1 Event Quiz Interface

Given a target event e.g. "changing a tire", a set of N temporally coherent microshots from K positive and negative videos (randomly sampled) are selected to populate the quiz interface. We ensure that each microshot within a set, are roughly spread across the length of the source video to capture as much diversity as possible. The temporal structure between the microshots within a video is also presented through progress timeline, to provide the player a basic idea of the approximate temporal location of the microshot in the "parent" video.

For each video (as shown as a separate row in Fig. 2), a human player is asked to determine the presence or absence of the event – *changing a tire* with minimal "clicks". In order to achieve this, all microshots are initially hidden (shown as grayed frames in Fig. 2) from the game player. On click, a spatially downsampled version (animated GIF sequences used to reduce the network overhead of the gameplay and low spatial resolution ensures that players cannot recognize faces/other minute details to memorize a particular class of event) of the microshot is revealed to the player. For every request made to reveal a microshot, a player looses a fixed number of points. Consequently, for all true positive or negative decisions made, points are retained whereas, for false positive or negative decisions, points are lost. This scoring system ensures rewarding correct identification of a video sample in minimal revelation requests.

Once the player is done with one quiz session, the system displays the final scores (based on the scoring rule described above) and the accuracy (number of correct decisions divided by $2K$). Note that the order in which microshots are requested for revelation from the player is also recorded (shown as numbers over each microshot in Fig. 2) as they help understand how humans accumulate the evidence in reaching the final judgment about event occurrence.

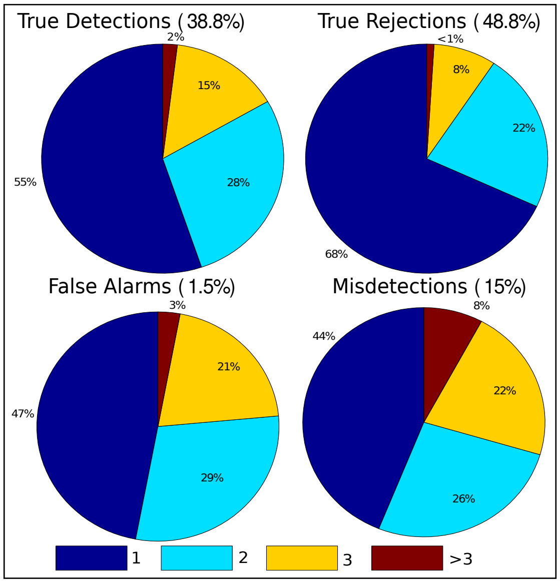


Figure 3: Human decision trends with number of microshots requested before deciding presence/absence of an event.

In the next section, we discuss some of the insights we have drawn from our user study over a fraction of the TRECVID MED 2013 ad-hoc events dataset.

3.2 Human Study

We used our event quiz interface to conduct an in-depth study of human performance over videos from 10 complex event categories. For this experiment, we recruited 8 human players of age between $25-35$, without no previous knowledge of the TRECVID MED (Multimedia Event Detection) Task. The dataset contains diverse videos from events including Tailgating, Beekeeping, Wedding Shower etc. All players were requested to play the event quiz game 2 times for each target event, generating about $1,600$ judgments. Also each event is viewed by multiple users for reliability. While playing the game, the players were refrained from using web search to find relevant visual examples of the target event, to solely judge the human understanding given only the textual definition of an event.

Fig. 3 provides analysis of human decision trends with respect to the number of microshots requested to reveal. The 4 pie-charts shown here correspond to different outcomes of user decisions against different number of microshots requested to reveal. The slices in the charts indicate the number of microshot revelation requests made. As evident, in approximately 87% of the test cases, humans make correct decisions about an event. Another surprising observation is that only one microshot is good enough in 55% of the test cases to identify an event correctly. Furthermore, in 68% of the cases where humans correctly rejected a video, required viewing just one single microshot. Thus, seeing one microshot provides more cues towards making true negative than true positive decisions. The pie-charts in the bottom of Fig. 3 show the failures made by humans.

Videos that induce confusion in the user, affect the false alarms and mis-detection trends, in which cases, more than one microshot revelations are being made. To develop an understanding on which class of events affect the human judgment, we provide an event specific breakdown of avg. accuracies for true detection and rejection cases in Fig. 4.

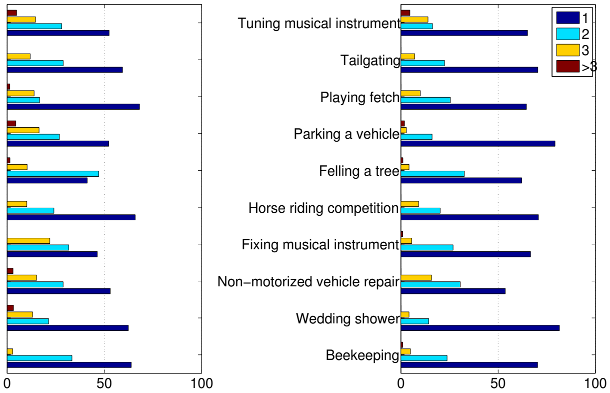


Figure 4: Event level break-down of Top 2 pie-charts from Fig. 3: The bar chart on the left shows the percentage of videos in each event category, correctly detected using different number of microshots revealed. The bar chart on the right, indicates the same for true rejections. The different number of microshots revealed are color coded as : blue (1), cyan (2), yellow (3), and brown (4) with numbers indicating # of microshots revealed.

We observe that for three specific events e.g. "Parking a Vehicle", "Tuning a musical instrument", "Non-motorized Vehicle Repair", and "Wedding Shower", humans need to look at more than 3 microshots for correct identification in relatively more cases ($\sim 4\%$ more) than other events. This can be attributed to the inherent complexity of these events. Intuitively, "Tuning" as opposed to "playing" or "repairing" a musical instrument can only be justified using relatively more evidences. Some events contain signature contextual evidences such as Horse/Jockey in "Horseriding Competition", Bee Farms/Beekeper wearing protecting clothing in "Beekeeping" etc. which have higher likelihood to be seen by humans, trivializing correct identification of these events. Hence just 1 microshot is sufficient, for these events to achieve high recognition performance. This observation provides a practical psychological insight, which we believe can be exploited to train specialized concept detectors.

Similar to the chart on left, a single microshot is sufficient for rejecting a large number of videos for most event categories. Thus, these microshots are ideally the MNE which can facilitate early rejection of videos in recognition context. A slightly more verbose feedback from the players, such as "I believe the video does not represent *Horseriding* because, (a) I See . . .", (b) I do not see . . .", etc. can be used to identify event specific concepts.

In other words, the MNEs discovered through the EQI, can be used to explain complex events more naturally. Specifically, these can further help us answer what evidence or concepts a human looks for to make conclusive judgments about an event. With some more sophistication, we may also be able to answer if there exists any explicit temporal relationship across evidence that are useful in the recognition process. The first an foremost step before investigating the above questions is to validate the credibility of the discovered MNEs for automated recognition. Hence we propose a simple video retrieval experiment to back up our hypothesis. The following section attempts to adapt the discovered MNEs into a basic retrieval framework.

4. RETRIEVAL

Traditional video retrieval techniques either exhaustively use all possible constituent shots from a video or automatically select discriminative [2] shots to form a query for retrieval. We are interested in investigating how MNEs can be efficiently used to formulate such queries. Thus, during a typical session, based on the order of the microshots revealed to a player, we identify microshots that are most representative towards making a positive judgment (selection or rejection of a video belonging to the target event category). For each microshot revealed, a separate query is generated, while the order of revelation associated with each microshot is later used to weigh the corresponding retrieved results.

Recall, N being the total number of microshots per video sample, available for revelation, lets consider Q being the number actually revealed to the player. Let i be the order of the revealed microshot, thus the higher the value of i, the latter it has been revealed to the human player. Consequently, it is likely to be more discriminative towards correct identification or rejection of a video with respect to a target event. We first represent each revealed microshot or query ($\mathbf{m}_i$) as a vector obtained from the bag-of-features representation (discussed in Section 4.1).

We perform a nearest neighbor search for each of the Q revealed microshots across a pre-indexed database of microshot vectors. Each of the P nearest neighbor to $\mathbf{m}_i$, say $\mathbf{x}_i^{(j)}$ in the set $X_i = \{\mathbf{x}_i^{(1)}, \mathbf{x}_i^{(2)}, \ldots, \mathbf{x}_i^{(P)}\}$ contributes a real valued vote ($v_i^{(j)}$, conforming to the set of votes V_i). The votes are computed based on the neighbors' distances to the revealed microshot $\mathbf{m}_i$, according to the following equation:

$$v_i^{(j)} = (N - Q + i) \times \exp(-|\mathbf{x}_i^{(j)} - \mathbf{m}_i|). \qquad (1)$$

The term $(N - Q + i)$ ensures a simple weighting mechanism penalizing nearest neighbors to those microshots that are less discriminative towards identifying a target event, and rewarding those that are revealed later to the player. Note that the above formulation fits perfectly in all cases, irrespective of the number of microshots revealed, while also capturing their order of revelation.

After we have obtained the nearest neighbors to each revealed microshot, our objective is to come up with a sorted list of videos (L), whose each entry is paired as $< Video, Vote >$. To perform this task, we employ a majority voting scheme listed in 1. The key idea of this algorithm is to add votes from neighbors that share their "parents" (recall microshots are sampled from a video).

Note that the algorithm 1 does not incorporate temporal relationship between multiple microshots used as queries, so we used a separate strategy to include this additional information. Instead of aggregating weights from all candidates, we only considered those candidates that are temporally consistent with the query. However, through experiments we found this additional information did not help improve retrieval performance and hence not reported.

We also leverage on the true negative decisions made by players in a setting discussed in Section 4.2. Player decisions that end up being false positives and negatives are not used in the retrieval process.

4.1 Microshot Representation

We use two complementary representations for each microshot. Recall that a microshot consists of M-consecutive frames, we use the center most frame to extract appearance based representation, while the entire microshot for motion based representation. SIFT [13] features are extracted from the center-most frame, and dense trajectory features described using motion boundary histograms [26] are extracted from the entire microshot, form the bases of the appearance and motion representation of a given microshot, respectively. A soft quantization based bag of visual words model is employed to create the final representations for each microshots. We use a protocol empirically similar to [26] to generate the optimal vocabulary in both cases.

These feature representations are independently used to construct pre-built indexing structures using a hierarchical k-means tree algorithm available in [19]. Note that, this step can be easily replaced with any algorithm capable of scalable enough to search in high-dimensional space.

Algorithm 1: Majority voting algorithm used in our video retrieval technique.

```
Procedure RetrieveVideos ($\{\mathbf{m}_1, \ldots \mathbf{m}_Q\}$, $N$, $P$)
Input: Set of queries $\{\mathbf{m}_1, \ldots \mathbf{m}_Q\}$,
Number of Microshots available per video ($N$),
Nearest Neighbors ($P$),
Database ($D$)
Output: Ranked List of videos ($L$)
$L \leftarrow \{\}$;
for each $\mathbf{m}_i \in \{\mathbf{m}_1 \ldots \mathbf{m}_Q\}$ do
    $X_i \leftarrow$ NNsearch($\mathbf{m}_i, D, P$);
    for each $\mathbf{x}_i^{(j)} \in X_i$ do
        // Calculate vote from $j$-th neighbor $\mathbf{x}_i^{(j)}$
        $v_i^{(j)} \leftarrow (N - Q + i) \times e^{(-|\mathbf{x}_i^{(j)} - \mathbf{m}_i|)}$;
    // Add neighbor votes from common "parents"
    $L_i \leftarrow$ addVotes($V_i, X_i$);
$L \leftarrow$ merge($L_i, L$);
$L \leftarrow$ sort($L$);
return (L);
```

4.2 Leveraging Negative Cues

In our human study, we observe that players can quickly reject some videos by viewing as few as a single microshot. In other words, given such shots, the players are certain that the video does not contain the target event . For example as shown in Figure 5, considering the event *parkour*, if a player sees an indoor scene, the entire sequence can be immediately rejected. On the other hand, if the microshot contains some evidence belonging to an outdoor scene with a handful of people, humans may need additional evidence to make the final decision.

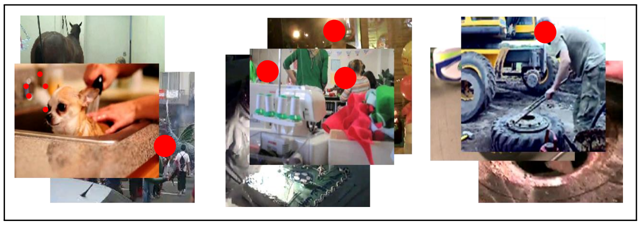


Figure 5: Representative frames from top 3 discovered negative microshot clusters for *Parkour*.

Motivated by the above fact, we propose to quickly reject the negative videos by learning from the user response. This can dramatically save the online testing time for event detection. Considering a single event,

- We first collect the video shots, such that after viewing this single shot, the user rejects the video. These video shots are considered to constrain strong cues against the event.

- We then cluster the above shots to S clusters. We assume that each cluster represents a strongly negative visual cue.

- For each cluster, we train a one-class SVM [16]. A one-class SVM can capture the visual properties of the clustering, without the need for defining the "positive" microshots.

- In the online test time, we first randomly sample 1 (or 2) microshots from each test video, and apply the S one-class SVMs. If any score of the S-classifiers is high, we will exclude the video for further processing.

In practice, we set $L = 5$, which can well capture the "negative clusters" based on visualization. Figure 5-6 show some top detected representative frames from the negative clusters for events *Birthday Party* and *Parkour*. We can see, for example microshots containing animals, indoor scenes, and fixing vehicles are strong negative cues for presence of *Parkour* event in a video.

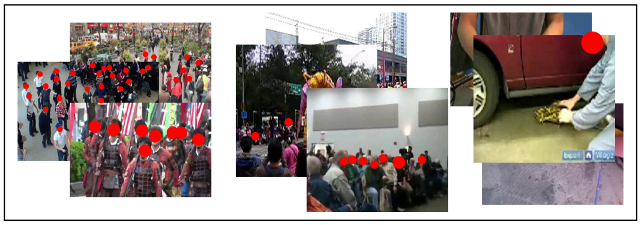


Figure 6: Representative frames from top 3 discovered negative microshot clusters for *Birthday Party*.

Recall the statistics presented in Fig. 3, human players can correctly reject negative samples of a target event by looking at one or atmost two microshots in $\sim 90\%$ cases. Thus, the online testing time can be dramatically reduced, as lots of videos are rejected without the time-consuming process of extracting features. In particular, even by performing simple exhaustive search without any sophisticated indexing structure employed at dataset level, we observed a $6-7$ times speed up in search time. Consequently, the memory allocated for loading a significant portion of the database (for efficient search) is reduced. The above approach is similar to performing a single-level cascade classification on temporal scale. A similar idea on the spatial scale has been widely used in object detection [24].

We explain our experimental protocol in the next section starting with the datasets we used for evaluation and discuss some interesting results thereafter.

5. EXPERIMENTS

In this paper, we are considering video-example based event retrieval. In other words, given a query video, we consider the videos in the database with same event label as positive. We use mAP to compare the performance of different methods. The following sections detail our extensive experiments on two widely used event recognition datasets released by NIST as part of TRECVID MED competition organized since 2010 [1].

5.1 Datasets

The first dataset is called the TRECVID MED 2013 Ad-hoc events dataset (referred as MED13 ADHOC) and has already been used to report our human study in Section 3.2. It consists of $1,497$ videos distributed over another 10 different event classes listed as follows: *(E031) Beekeeping*, *(E032) Wedding shower*,*(E033) Non-motorized Vehicle repair*, *(E034) Fixing musical instrument*, *(E035) Horse riding competition*, *(E036) Felling a tree*, *(E037) Parking a vehicle*, *(E038) Playing fetch*, *(E039) Tailgating*, *(E040) Tuning musical instrument*.

The second one is a combination of complex events from the event collections released in 2011 and 2012, hereafter referred as MED TEST. This dataset consists of $3,489$ videos from 20 different complex event categories. These are: *(E006) Birthday party*, *(E007) Changing a vehicle tire*, *(E008) Flash mob gathering*, *(E009) Getting a vehicle unstuck*,*(E010) Grooming an animal*, *(E011) Making a sandwich*, *(E012) Parade*, *(E013) Parkour*, *(E014) Repairing an appliance*, *(E015) Working on a sewing project*, *(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*, and *(E030) Working on a metal crafts project*.

All videos in the datasets are approximately uniformly distributed over all event classes, and are typically recorded by amateur consumers approximately at 30 fps with no specific resolution, under unconstrained scenarios. Also videos from these events have large degree of intra-class visual variance (e.g. *Attempting a Board trick* refers to both *Snow boarding* and *Skate boarding*), and in many cases demonstrate subtle inter-class visual variance (e.g. *Cleaning an Appliance* and *Repairing an appliance*).

5.2 Implementation

We implement the proposed Event Quiz Interface in a web based framework. A screen capture of a typical query session is shown in Fig. 7. The camera icon with text overlay "Show Microshot" is part of the actual game interface design. A player initially sees a set of camera icons, and attempt to judge whether a video contains the specific event (in this case "parade") by requesting to reveal a minimal number of microshots. Such minimal set of revealed microshots is deemed the MNEs for event recognition. Players are given incentives in the form of points (indicated in the rightmost column) for correct identification of an event by seeing minimal evidence.

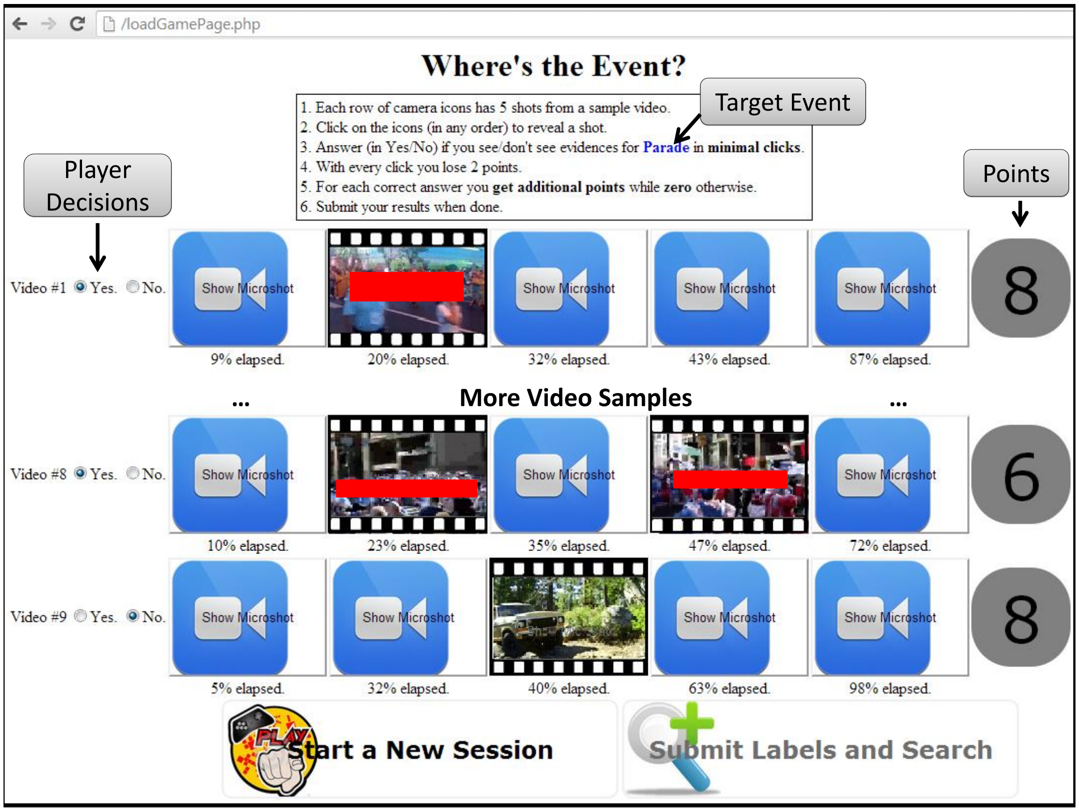
The versatility of the interface allows us to use it in testing mode, where queries can be formulated and simultaneously be used for retrieval purpose (*Submit labels and Search* button in Fig. 7). Before each session, EQI automatically generates a random target event name from either of MED13 ADHOC or MED TEST datasets, and initializes the placeholders (camera icons in Fig. 7) with subsampled animated images of the microshots. K is set to 5, indicating the number of positive and negative samples to be used per session. In practice, we find that smaller K cannot capture the contents of the video, yet larger K will create more burdens for the user.

Each microshot is pre-sampled from its parent video subjected to an interestingness threshold (I_δ). Only microshots with sufficient interestingness (in terms of appearance and dynamics) are used. For example, dark or static microshots will not be selected. The interestingness (I) of a microshot of length M is computed as a weighted function of entropy content with respect to appearance and motion:

$$I = -\alpha \sum_{i=1}^{M} P_a(i)\log[P_a(i)] - \beta \sum_{i=1}^{M} P_m(i)\log[P_m(i)], \qquad (2)$$

where $P_a(i)$ and $P_m(i)$ are the probability mass functions for ap-

[1] http://www.nist.gov/itl/iad/mig/med.cfm



Where's the Event?

1. Each row of camera icons has 5 shots from a sample video.
2. Click on the icons (in any order) to reveal a shot.
3. Answer (in Yes/No) if you see/don't see evidences for Parade in **minimal clicks**.
4. With every click you lose 2 points.
5. For each correct answer you **get additional points** while **zero** otherwise.
6. Submit your results when done.

Video #1 ● Yes. ○ No.

Video #8 ● Yes. ○ No.

Video #9 ○ Yes. ● No.

Figure 7: Screen capture of a typical game play session on the event quiz web interface using the MED13 ADHOC dataset, with *Parade* being used as a target event.

pearance and motion, respectively. $P_a(i)$ is directly derived from horizontal and vertical intensity gradients and P_m(i) is computed using successive frame difference for ith frame in a candidate microshot. More sophisticated methods [2, 11] can be applied to select better microshots which would involve sacrificing computational speed, and hence not considered within the purview of this work.

During microshot extraction, motion entropy content of a sequence is weighed higher than that of appearance, as we are more interested in videos that are rich in action. This is ensured by empirically setting $\alpha = 0.3$, $\beta = 0.6$ and $I_\delta = 1$. Two types of features are extracted from the microshots created using the method described in Section 4.1. We set $M = 41$ as the microshot length (most human actions that are repetitive in nature such as jumping, running, walking, waving, clapping etc. can be discerned after watching 1.5s of a video footage captured under 30fps [6, 12]), thus the twenty first frame is marked as key frame. Dense Trajectory Features [26] extracted from a microshot, and Dense SIFT features [13] extracted from the corresponding keyframe, are quantized into two separate vocabularies of $5,000$ and $2,000$ respectively. These vocabularies are used subsequently to derive the bag of features representations corresponding to motion and appearance respectively.

For every microshot used in the query process, we obtain 128 nearest neighbors using a fast approximate nearest neighbor search as discussed in Section 4. For all experiments, precision at a range of top k-ranked retrieved results (k = 5,10, 15, 20, 30, 40, 60,80, 100 are computed before returning the final mean Average Precision.

5.3 Comparisons

Although the MED TEST dataset used in the paper is used in previous event classification experiments in [2,11,12,23], the MED13 ADHOC dataset is brand new. Additionally, since our focus is on the retrieval aspect of videos containing complex events, it is extremely difficult to compare our performance with the previous methods that perform classification and report results on MED TEST dataset.

However, to demonstrate the performance of our method, we compare with two different baseline methods. The first one does not perform any smart selection of evidence before retrieval (BL-A). In other words, this baseline uses all microshots available per query video exhaustively. We implement another baseline method that performs automatic discriminative microshot selection based on the technique proposed in [2]. In this baseline, we cluster all microshots from a video to a few representative ones. The number of representative shots are empirically determined to be 5. This is referred as BL-B in the next few sections.

6. RESULTS

We take the opportunity to report our results on the MED13 ADHOC dataset here. A summary of experiments under this setting is provided in Tab. 1. The number of queries used for retrieval is listed in column 2. This is followed by the average precisions obtained using two baselines: all microshots (BL-A) and automatically selected microshots using clustering (BL-B). The next two columns report the respective APs with MNEs selected by human players without and with quick rejection scheme discussed in Section 4.2. In all cases appearance information is used to perform retrieval.

Events	BL-A	BL-B	MNE	MNE+QR
Beekeeping	3.47	4.12	20.96	20.86
Wedding shower	2.87	2.05	17.23	17.42
Non-motorized Vehicle repair	2.56	3.35	16.90	17.09
Fixing musical instrument	3.52	3.09	19.26	19.69
Horse riding competition	4.60	5.21	21.46	11.91
Felling a tree	5.47	5.25	20.86	11.27
Parking a vehicle	3.09	6.11	17.04	17.35
Playing fetch	2.73	4.08	16.62	16.74
Tailgating	1.75	3.15	15.48	14.97
Tuning musical instrument	3.95	4.06	18.26	18.56
Mean Average Precision	3.41	4.07	18.47	18.59

Table 1: Average Precision (%) for all events from TRECVID MED 2013 Ad-hoc Event collection data using different retrieval methods.

Our experiments back up our hypothesis that video retrieval with MNEs achieve around 14% absolute improvement in AP over that obtained using a state-of-the-art method (similar to [2]) that uses automatic evidence selection. There is slight improvement in AP when early rejection of irrelevant videos based on negative cues is employed.

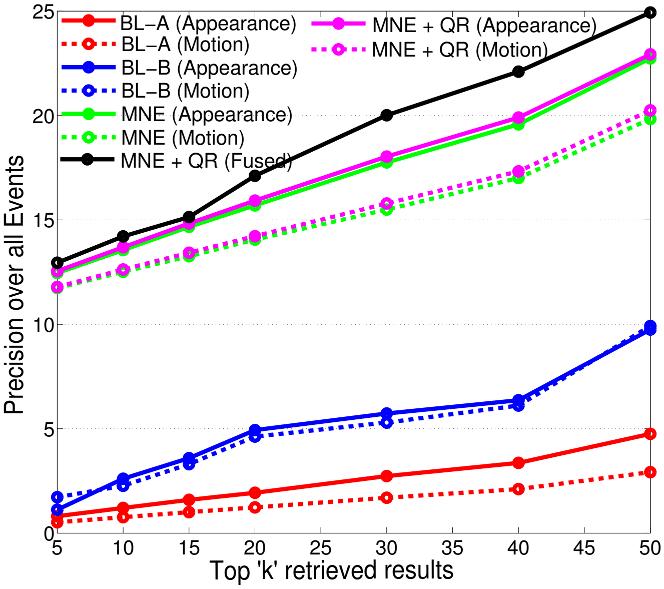


Figure 8: Summary of retrieval on MED13 Ad-hoc event dataset.

Fig. 8 offers a better understanding of the retrieval behavior on this dataset using different methods tested with both the appearance and motion based representations. We report the precision of retrieval with respect to top $k = 5$, 10, 15, 20, 30, and 50 videos returned. For legibility, performance on retrieval systems that are using appearance based representations are indicated in solid lines whereas the same using motion are shown using dashed lines. The retrieval performance using all available evidence (BL-A) for a query video is shown in red, automatically selected evidence based on [2] (BL-B) in blue, Minimally Needed Evidences (MNE) in green and quick rejection from negative cues paired with Minimally Needed Evidences (MNE + QR) in magenta. Finally, results based on early fusion of both motion and appearance representations for MNE+QR is indicated in solid black line.

According to our observation, appearance based representation yields slightly better performance than the motion based one for all methods, however fusion of both boosts the overall performance. This reinstates that both representations carry complementary information.

Events	[17]	MNE+QR (Fused)
Birthday party	15.9	35.6
Changing vehicle tire	11.8	23.4
Flash mob	30.5	34.3
Vehicle unstuck	15.9	28.2
Grooming animal	21.4	28.5
Making sandwich	13.8	31.6
Parade	22.1	34.8
Parkour	21.9	37.1
Repairing appliance	22.1	32.1
Sewing project	10.1	24.4
Bike trick	11.0	21.3
Cleaning appliance	7.2	19.1
Dog show	13.0	24.5
Giving directions	11.4	22.9
Marriage proposal	2.5	16.3
Renovating home	20.7	29.1
Rock climbing	6.5	18.5
Town hall meeting	5.5	12.3
Winning race w/o vehicle	8.5	18.6
Metal crafts project	3.0	16.2
Mean average precision	13.1	25.4

Table 2: Average Precision (%) over all 20 events from TRECVID MED 2011-12 Event collection data.

In Tab. 2, we report our performance on the MED TEST dataset consisting of 20 events. We also compare our retrieval performance against a state of the art technique published in [17]. For clarity, we only indicate the respective average precision per event obtained using our best performing method (MNE+QR on fused motion and appearance representations). Although, the representation used in [17] cannot be directly compared with the proposed method in this paper, it gives a general overview of the complexity of retrieval in the MED TEST dataset. It is interesting to note that a simple representation coupled with an equally simple retrieval technique can report 12% improvement over the performance published by a state of the art method that uses a more semantically sophisticated representation, using only *careful selection of evidence*.

Finally, in Fig. 9 we provide some qualitative results visualizing the MNEs discovered by our event quiz interface from three MED TEST events. Each block of filmstrips shown in Fig. 9 shows MNEs from 3 video samples belonging to an event category. Each film strip shows randomly sampled, temporally coherent evidence - with MNEs selectively shown. Each video can have multiple MNEs as they are seen by multiple human players. The blue rectangular bars underneath each film strip, together with overlaid red bars, indicate a timeline and approximate temporal locations of the MNEs within the timeline, respectively. We conjecture that this information can be applied as an additional temporal prior while looking for evidence.

7. CONCLUSION & FUTURE WORK

In this paper, we propose a new framework of human cognitive capability in recognizing complex events in videos. We hypothesize and validate that humans can recognize a complex event by viewing just a very small number of microshots. We conceived the *Event Quiz Interface (EQI)*, an interactive tool mimicking some of

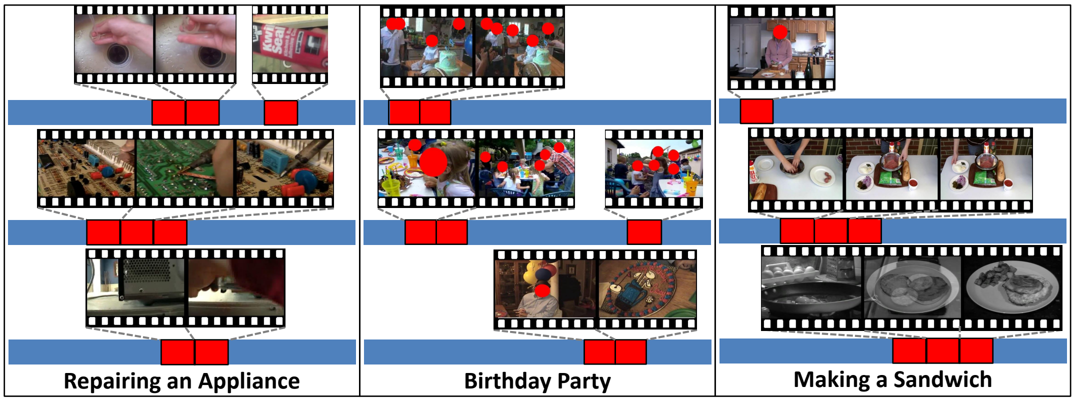


Figure 9: Qualitative depiction of MNEs for three events from the MED TEST dataset. Each red segment corresponds to a microshot revealed by the user.

the key strategies employed by the human cognition system to develop high-level understanding of complex events in unconstrained web videos. We also introduced the notion of *Minimally Needed Evidence (MNE)* to predict the presence or absence of an event in a video. We performed conclusive experiments to demonstrate significant improvement in event retrieval performance on two challenging datasets released under TRECVID using off-the-shelf feature extraction techniques applied on MNEs versus evidence collected sequentially.

We are currently investigating how our proposed EQI can be extended to incorporate human feedback in a more organic way. Simple free-text inputs to describe the evidence observed while making a decision, can pave a better way to perform annotation, and consequently obtain refined training samples for supervised learning of concepts. We believe, as we ingest more event samples into the EQI, we can generate a consensus on the kind of microshots that are helpful for large scale event detection.

Acknowledgments: This work was supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center contract number D11PC20070. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either express or implied, of IARPA, DoI/NBC, or the U.S. Government.

8. REFERENCES

[1] L. Cao et al. IBM Research and Columbia University TRECVID-2011 Multimedia Event Detection (MED) System. In *NIST TRECVID Workshop*, Gaithersburg, MD, December 2011.

[2] L. Cao, Y. Mu, A. Natsev, S.-F. Chang, G. Hua, and J. R. Smith. Scene aligned pooling for complex video recognition. In *ECCV*, 2012.

[3] M.-S. Dao, G. Boato, and F. G. B. DeNatale. Discovering inherent event taxonomies from social media collections. In *ICMR*, pages 48:1–48:8, 2012.

[4] J. Deng, J. Krause, and L. Fei-Fei. Fine-grained crowdsourcing for fine-grained recognition. 2013.

[5] D. Doermann and D. Mihalcik. Tools and techniques for video performance evaluation. In *IEEE ICPR*, volume 4, pages 167–170, 2000.

[6] M. Hoai and F. De la Torre. Max-margin early event detectors. In *IEEE CVPR*, pages 2863–2870, 2012.

[7] P. Isola, J. Xiao, A. Torralba, and A. Oliva. What makes an image memorable? In *IEEE CVPR*, pages 145–152, 2011.

[8] Y.-G. Jiang, S. Bhattacharya, S.-F. Chang, and M. Shah. High-level event recognition in unconstrained videos. *International Journal of Multimedia Information Retrieval*, 2(2):73–101, 2013.

[9] 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*, Gaithersburg, MD, November 2010.

[10] T. Judd, K. Ehinger, F. Durand, and A. Torralba. Learning to predict where humans look. In *IEEE ICCV*, pages 2106–2113, 2009.

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

[12] W. Li, Q. Yu, H. S. Sawhney, and N. Vasconcelos. Recognizing activities via bag of words for attribute dynamics. In *IEEE CVPR*, pages 2587–2594, 2013.

[13] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60:91–110, 2004.

[14] Y.-F. Ma, L. Lu, H.-J. Zhang, and M. Li. A user attention model for video summarization. In *ACM MM*, pages 533–542, 2002.

[15] V. Mahadevan and N. Vasconcelos. Spatiotemporal saliency in dynamic scenes. *IEEE TPAMI*, 32(1):171–177, 2010.

[16] L. M. Manevitz and M. Yousef. One-class svms for document classification. *Journal of Machine Learning Research*, 2:139–154, 2002.

[17] M. Mazloom, A. Habibian, and C. G. Snoek. Querying for video events by semantic signatures from few examples. In *ACM MM*, MM '13, pages 609–612, 2013.

[18] K. McGuinness et al. The AXES PRO video search system. In *ICMR*, pages 307–308, 2013.

[19] M. Muja and D. G. Lowe. Fast approximate nearest neighbors with automatic algorithm configuration. In *VISAPP*, pages 331–340, 2009.

[20] P. Natarajan et al. BBN VISER TRECVID 2011 Multimedia Event Detection System. In *NIST TRECVID Workshop*, December 2011.

[21] C. G. M. Snoek et al. MediaMill at TRECVID 2013: Searching concepts, objects, instances and events in video. In *NIST TRECVID Workshop*, November 2013.

[22] C. G. M. Snoek, M. Worring, D. Koelma, and A. Smeulders. A learned lexicon-driven paradigm for interactive video retrieval. *IEEE Trans. Multimedia*, 9(2):280–292, 2007.

[23] K. Tang, L. Fei-Fei, and D. Koller. Learning latent temporal structure for complex event detection. In *IEEE CVPR*, pages 1250–1257, 2012.

[24] P. A. Viola and M. J. Jones. Rapid object detection using a boosted cascade of simple features. In *IEEE CVPR*, pages 511–518, 2001.

[25] C. Vondrick and D. Ramanan. Video Annotation and Tracking with Active Learning. In *Neural Information Processing Systems*, 2011.

[26] H. Wang, A. Kläser, C. Schmid, and L. Cheng-Lin. Action Recognition by Dense Trajectories. pages 3169–3176, June 2011.

[27] M. Wang and X.-S. Hua. Active learning in multimedia annotation and retrieval: A survey. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(2):10, 2011.

[28] Z. Wang, M. Zhao, Y. Song, S. Kumar, and B. Li. YouTubeCat: Learning to categorize wild web videos. In *IEEE CVPR*, 2010.

[29] J. Yuen, B. C. Russell, C. Liu, and A. Torralba. LabelMe video: Building a video database with human annotations. In *IEEE ICCV*, pages 1451–1458, 2009.

[30] E. Zavesky and S.-F. Chang. CuZero: Embracing the Frontier of Interactive Visual Search for Informed Users. In *ACM MIR*, Vancouver, British Columbia, Canada, October 2008.

[31] G. Zhu, M. Yang, K. Yu, W. Xu, and Y. Gong. Detecting video events based on action recognition in complex scenes using spatio-temporal descriptor. In *ACM MM*, pages 165–174, 2009.