# ISOMER: Informative Segment Observations for Multimedia Event Recounting

Chen Sun*, Brian Burns‡, Ram Nevatia*, Cees Snoek†, Bob Bolles‡, Greg Myers‡,
Wen Wang‡ and Eric Yeh‡

*University of Southern California
{chensun|nevatia}@usc.edu

‡ SRI International
{burns|bolles}@ai.sri.com
{gregory.myers|eric.yeh}@sri.com
wwang@speech.sri.com

†University of Amsterdam
cgmsnoek@uva.nl

## ABSTRACT

This paper describes a system for multimedia event detection and recounting. The goal is to detect a high level event class in unconstrained web videos and generate event oriented summarization for display to users. For this purpose, we detect informative segments and collect observations for them, leading to our ISOMER system. We combine a large collection of both low level and semantic level visual and audio features for event detection. For event recounting, we propose a novel approach to identify event oriented discriminative video segments and their descriptions with a linear SVM event classifier. User friendly concepts including objects, actions, scenes, speech and optical character recognition are used in generating descriptions. We also develop several mapping and filtering strategies to cope with noisy concept detectors. Our system performed competitively in the TRECVID 2013 Multimedia Event Detection task with near 100,000 videos and was the highest performer in TRECVID 2013 Multimedia Event Recounting task.

## Categories and Subject Descriptors

I.2.10 [**Artificial Intelligence**]: Vision and Scene Understanding—*Video Analysis*

## General Terms

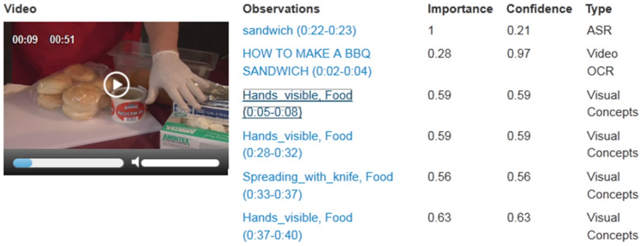
Algorithms, Experimentation

## Keywords

video, event classification, event recounting

## 1. INTRODUCTION

The amount of video data generated and stored from institutional and user sources is increasing at an amazing speed.



| Observations | Importance | Confidence | Type |
|---|---|---|---|
| sandwich (0:22-0:23) | 1 | 0.21 | ASR |
| HOW TO MAKE A BBQ SANDWICH (0:02-0:04) | 0.28 | 0.97 | Video OCR |
| Hands_visible, Food (0:05-0:08) | 0.59 | 0.59 | Visual Concepts |
| Hands_visible, Food (0:28-0:32) | 0.59 | 0.59 | Visual Concepts |
| Spreading_with_knife, Food (0:33-0:37) | 0.56 | 0.56 | Visual Concepts |
| Hands_visible, Food (0:37-0:40) | 0.63 | 0.63 | Visual Concepts |

**Figure 1: Our system's recounting results for a *making a sandwich* video**

Efficient management and utilization of this data requires it to be automatically annotated and summarized in meaningful ways. These annotations can be at various levels of detail, including a label for the overall theme but also for its important, critical components. Take for example a user generated video of a child's *birthday party*. We can label the entire video with an event label of *birthday party*. We can also describe the video in more detail by isolating its important segments which could, for example be, instances of *singing*, *blowing out candles* and *cutting the cake*. An ordered collection of segments containing such actions may be useful in browsing the video in a summary form and in verifying whether the video is in fact of a *birthday party*. Finally, labeling the segments by their actions may assist in retrieval of videos based on some combination of actions that are not necessarily specified in advance. In this paper, we present techniques for video classification, summarization and generating textual descriptions.

Though our techniques are general, we focus on collections of user generated datasets, such as what we may find on YouTube. We show results of experiments on a large collection compiled by NIST (National Institute of Standards and Technology) as part of their TRECVID Multimedia Event Detection (MED) and Multimedia Event Recounting (MER) evaluations [13]. This dataset consists of over 100,000 videos comprising of 25 event classes and unrelated background videos. Videos may also contain audio, speech and text information. NIST has conducted rigorous evaluations on results submitted by various participants on this dataset so good comparative data is available.

The purpose of the MED evaluation was to character-

*ICMR* '14, Apr 01-04 2014, Glasgow, United Kingdom
Copyright 2014 ACM 978-1-4503-2782-4/14/04 ...$15.00.

ize the performance of multimedia event detection systems, which aim to detect user-defined events involving people in massive, continuously growing video collections, such as those found on the Internet. This is an extremely challenging problem, because the contents of the videos in these collections are completely unconstrained, and the collections include varying qualities of user-generated videos, which are often made with hand-held cameras and have jerky motions and wildly varying fields of view.

The goal of multimedia event recounting is to give users a human-understandable recounting for each clip that the MED system deems to be positive. In Figure 1, a *making a sandwich* video's recounting has speech information (*sandwich*), text information and object/action information. Providing such evidence is not so straightforward, because humans usually think of an event in terms of specific associated semantic concepts, but the reliability of detectors for most individual semantic concepts is poor. The purpose of the MER evaluation was to assess the quality of recounting evidence associated with the MED retrieval results.

Figure 2 illustrates an overview of our system. Our approach is to use a variety of modalities, including visual, audio and text. We infer action, object and word level concepts at segment level from low level features and combine them at the video level to provide event label likelihoods. For each event label, we find *informative segments* and use observations derived from them to create MER output; we name this approach to be ISOMER. ISOMER employs concept mapping and filtering strategies to cope with noisy detectors, and offers a framework to select and merge observations from different modalities. There has been significant work on the MED task in the last 2 to 3 years but MER is a new task, especially for unconstrained, user generated videos.

The rest of this paper is organized as follows: Section 2 describes the related work on MED and MER tasks. Section 3 describes our event classification framework for MED task, and its performance in the NIST TRECVID 2013 MED evaluation. Section 4 described our event recounting framework for MER task, and its performance in the TRECVID 2013 MER evaluation.

## 2. RELATED WORK

Low level features have been applied to multimedia event detection, such as SIFT [9] for static frames and STIP [7] for video. Recently, dense trajectory features [21] were shown to have good performance on challenging datasets. Its feature vectors were obtained by tracking densely sampled points and describing the volume around tracklets by optical flow, gradient and motion boundary features. To aggregate video level feature vectors, [18] achieved good performance by applying Fisher Vector coding on low level features.

Multimedia event recounting is a type of summarization. It differs from most current summarization efforts [19] in being specific to an event class: it is a selection of video segments and extracted content relevant to the event. In [3] , this was called topic-oriented multimedia summarization, and was generated by selecting a set of event-relevant semantic concepts before the video was processed and then localizing these concepts in the video using the peaks of their individual detection scores. This process was also used in [12].

Detected semantic concepts (objects, scenes, actions) are some of the most important observations to report while recounting. However, their detection in unconstrained video is still a challenge, and videos that are part of the same event class may contain very different subsets of the event-relevant concepts. In [8], the latter problem was tackled by analyzing the contribution of each concept detection score to the overall event classification of the video. The event classifier used the concept detection scores as features, and the contributions were assessed from the terms summed to perform the classification. This method produced superior recounting results in the TRECVID 2012 recounting task; however, it still relied on the individual, noisy concept scores to localize the informative segments in the video. The method discussed in this paper is an advance on this idea, as it integrates the event classification and video segment localization steps into one process. This minimizes the dependence on individual concept detections for event-relevant summarization, producing the best results in the more recent TRECVID 2013 recounting task.

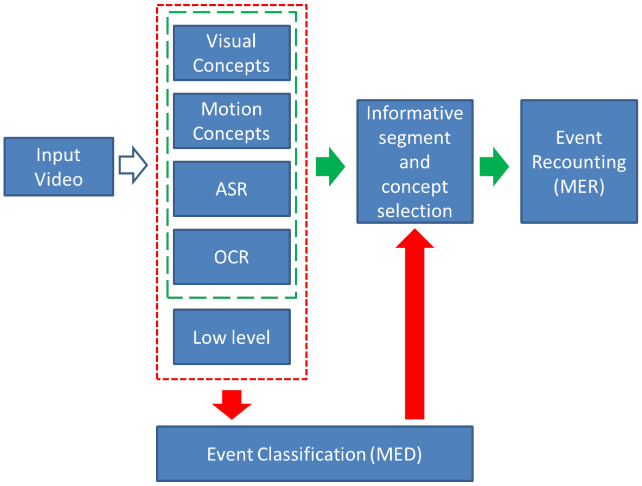


**Figure 2: An overview of our ISOMER system framework (best viewed in color). We extract visual and motion concepts features, ASR and OCR detections, as well as low level features. All of them (in the red box) are used to train event classifiers for MED task. After event label has been assigned, we use concept response features (in the green box) to select informative video segments and their concept based descriptions. These selected observations from different modalities are fused into a single video level event recounting.**

## 3. EVENT CLASSIFICATION

This section describes our approach to the classification of an entire video and provides evaluation results.

### 3.1 Video Features

We extract a comprehensive set of heterogeneous low-level visual, audio, and motion features; high-level semantic concepts for visual objects, scenes, persons, and actions; and semantic concepts from the results of automatic speech recognition (ASR) and video optical character recognition (OCR). We process each modality independently and then fuse the results.

### 3.1.1 Static Visual Features and Concepts

For the visual features and concepts, we relied on one event classifier based on low-level visual features and two event classifiers based on semantic features obtained from visual concept detector scores. Before we detail the event classifiers, we first detail their low-level and semantic features.

**Low-level features.** We extracted low-level visual features for two frames per second from each video. We followed the bag-of-codes approach, which considers spatial sampling of points of interest, visual description of those points, and encoding of the descriptors into visual codes. For point sampling, we rely on dense sampling, with an interval distance of six pixels and sampled at multiple scales. We used a spatial pyramid of 1x1 and 1x3 regions in our experiments. We used a mixture of SIFT, TSIFT, and C-SIFT descriptors [20]. We computed the descriptors around points obtained from dense sampling, and reduced them all to 80 dimensions with principal component analysis. We encoded the color descriptors with the aid of difference coding, using Fisher vectors with a Gaussian Mixture Model (GMM) codebook of 256 elements [14]. For efficient storage, we performed product quantization [6] on the features.

**Semantic features.** We detected semantic concepts for each frame using the low-level visual features per frame as input representation. We followed the approach in [4]. Our pool of detectors used the human-annotated training data from two publicly available resources: the TRECVID 2012 Semantic Indexing task [15] and the ImageNet Large-Scale Visual Recognition Challenge 2011 [2]. The former has annotations for 346 semantic concepts on 400,000 key frames from web videos. The latter has annotations for 1,000 semantic concepts on 1,300,000 photos. The categories are quite diverse and include concepts of various types; i.e. objects like *helicopter* and *harmonica*, scenes like *kitchen* and *hospital*, and actions like *greeting* and *swimming*. Leveraging the annotated data available in these datasets, together with a linear support vector machine (SVM), we trained 1,346 concept detectors in total. We then applied all available concept detectors to the extracted frames. After we concatenated the detector outputs, each frame was represented by a concept vector.

To arrive at a video-level representation for the low-level visual event classifier, we relied on simple averaging. We trained the classifier with a linear kernel SVM. To handle imbalance in the number of positive versus negative training examples, we fixed the weights of the positive and negative classes by estimating the prior probabilities of the classes on training data. For the two video event classifiers based on semantic features, we aggregated the concept vectors per frame into a video-level representation. On top of both concept representations per video, we used a non-linear SVM with $\chi^2$ kernel with the same fixed weights to balance positive and negative classes.

### 3.1.2 Motion Features and Concepts

We used two basic low-level motion features: Dense Trajectories (DTs) [21] and MoSIFT [1]. The raw features were encoded using first- and second-order Fisher Vector descriptors with a two-level spatial pyramid [18]. Descriptors were aggregated across each video.

Two event classifiers were generated based on action concept detectors. There are 96 action concepts annotated on the MED11 Event Kit provided by Sarnoff & UCF [5], and 101 action concepts from UCF 101 [16]. Sarnoff concepts contain actions that happen directly in MED videos, such as *throwing*, *kissing*, and *animals eating*. UCF 101 concepts are not directly relevant to MED videos: for example, *cliff diving* and *playing violin*. Nonetheless, including these concepts still improves event detection scores. The action concept detectors were trained based on DT features with linear SVMs, applied to small segments of videos and encoded by Hidden Markov Model Fisher Vector descriptors [17].

SVM with Gaussian kernel performed the classification task, whose parameters were selected using five-fold cross validation on the training set.

### 3.1.3 Audio Features

For our audio features, we extracted Mel-frequency cepstral coefficients (MFCCs) over a 10 ms window. MFCCs describe the spectral shape of audio. The derivatives of the MFCCs ($\delta$ MFCC) and the second derivatives ($\delta\delta$ MFCC) were also computed. The MFCC features were difference-coded with Fisher vectors using a 1024-element Gaussian Mixture Model. For classification, we used a linear kernel SVM.

### 3.1.4 ASR

To extract words from audio information, we ran an English ASR model trained on conversational telephone data and adapted to meetings data. We performed supervised and unsupervised adaptation for acoustic and language models. We used ASR to compute probabilistic word lattices, from which we extracted video-based one-gram word counts for MED, and local counts over 1 second intervals for MER. We then performed stemming to reduce the vocabulary size to about 40,000 words. The stemmed counts were mapped to features using a log-mapping.

For each event, we used those features to train a linear SVM with an $l$-1 penalty. The non-linear mapping was a sigmoid that was tuned empirically. The event profiles were obtained from the event kit text by using term frequency-inverse document frequency (TF-IDF) weightings to rank the relevance of non-stopwords.

### 3.1.5 OCR

We used SRI video OCR software to detect and recognize text appearing in MED13 video imagery [10]. This software recognizes both overlay text, such as captions that appear on broadcast news programs, and in-scene text on signs or vehicles. It was configured to recognize English language text.

After text recognition, we filtered the recognized text by its confidence score, retaining only those at 90% or greater. Because each line of video text was recognized independently, independent detections were grouped together into a single phrase if the amount of time between the two pieces of recognized text was less than 30 ms. These combined detections were used for the MER task.

For each event, we trained a log-linear classifier regularized with dropout over the frequency histogram of words from the preprocessed video OCR text identified in the training videos. In addition to the preprocessing, we also removed stopwords. At event detection time, we used the likelihood

| System | mean Average Precision |
|---|---|
| Static | 0.350 |
| Motion | 0.292 |
| Audio | 0.094 |
| ASR | 0.090 |
| OCR | 0.044 |
| Fused | **0.425** |
| Random | 0.003 |

**Table 1: MED performance (measured in mean average precision) on the MEDTest dataset**

of the video being a positive for the event, given the recognized keywords, as the event detection score.

### *3.1.6 Fusion*

We applied a simple late fusion method to combine the results from the various modalities: arithmetic mean. The detection scores were normalized using z score by removing the mean and scaling by the standard deviation. The normalization parameters were learned from the distribution of scores on the training set via cross-validation.

## 3.2 Performance Evaluation

TRECVID 2013 MED dataset has several partitions. For training, we use Event Kit 100EX dataset, which has 25 event classes and around 100 positive examples for each event. We also use Event Kit Background dataset, which has 4992 negative videos unrelated to all events. Performance of our system was assessed on the MEDTest dataset, it contains 23,542 videos.

Table 1 shows MED performance on the MEDTest dataset for six system configurations: Static only (consists of static visual features and concepts), Motion only (consists of motion features and action concepts), non-ASR audio only, ASR only, OCR only, and the full system (i.e., with all subsystems). We also report the expected mean AP by randomly assigning event classes.

The table clearly shows that all of the MED systems outperformed a random system. It also shows that the MED performance was dominated by the classifiers that exploited the visual content (static and motion). However, by using late fusion technique, the other modalities can still improve the overall performance [11].

# 4. EVENT RECOUNTING

The recounting consists of a concise textual summary of each piece of evidence and the source of the evidence, which may include action concepts, object and scene concepts, visible text, speech and non-speech sound patterns. For each piece of evidence, the recounting also includes a confidence score, an importance score indicating how important the evidence was in detecting the event, and spatial and temporal locations in the video where the evidence occurs.

## 4.1 MER Framework

Our system generates event recountings based on semantic concepts from the following multimedia sources: ASR, video OCR, and the 1,543 automatically detected visual objects, scenes, persons, and actions. We will first describe our methods for generating MER observations in each modality, and then our process for selecting and merging them.

### *4.1.1 Visual Observations*

As discussed in Section 3.1.1 and Section 3.1.2, by applying pre-trained visual and action concept detectors, each video was represented as a sequence of 1,346 object and scene concept responses, and 197 action concept responses.

However, there are two major problems to generate visual observations with the concepts: the first is that many of the concepts are not related to the high level events or are not discriminative enough for users; the second is the inherent uncertainty of visual and action concept detectors. We tackled these two problems by utilizing event prior information obtained from MED stage, and locating informative video segments and visual & motion concepts with the help of a linear SVM.

**Alignment of Concept Responses.** Before generating visual observations, we first aligned the detector responses of visual and motion concepts. Visual concepts were detected on static frames every half second, while motion concepts were detected on 100 continuous frames with a step size of 50 frames. We maintained the segment length of 100-frame, and used average pooling for visual concept responses within each segment, given by

$$\frac{1}{|C|}\sum_{c\in C}\mathbf{X}_c \tag{1}$$

where $\mathbf{X}_c$ is a concept response vector, $C$ is the set of frames with concept responses which fall in a certain 100-frame segment.

The pooled visual concept response vector was used with the action concept response vector to represent a video segment.

**Informative Segment Localization.** We applied discriminative event models to locate video segments and rank visual and motion concepts. It is assumed that each video has already been assigned an event class by the MED system.

For each event class $E$, a one-vs-rest linear SVM with weights $w_E$ was trained based on video level concept responses from the training dataset. A video level concept response was obtained by average pooling of responses over all video clips of the same video as defined in Equation 1.

In order to get a subset of video segments which is the most informative for event $E$, we applied the linear SVM of event $E$ to all segments with concept response vectors, and selected the top $K$ segments with highest SVM scores. We converted SVM scores into probabilities by the logistic regression function

$$P_E(c) = \frac{1}{1+\exp(-w_E\mathbf{X}_c)} \tag{2}$$

where $\mathbf{X}_c$ is clip $c$'s concept response.

$P_E(c)$ was used as the confidence and importance scores for video segment $c$. We then used a threshold to remove the unconfident segments.

Since each dimension of $\mathbf{X}_c$ reflects the confidence of a visual or action concept detector, we computed the element-wise product of $w_E$ and $\mathbf{X}_c$ and picked the top $D$ concepts with highest values. The selected concepts served as the description for the video segment.

| UCF 101 name | Mapped name |
|---|---|
| long jump | horizontal body movement |
| apply lipstick | hand movement |
| walking with dog | person walking |
| salsa spin | person dancing |

**Table 2: Some mapping examples for action concepts of UCF 101**

By using event specific weights to select concepts, we suppressed the noise of concept detectors, and removed the concepts that are common to most events.

**Concept Mapping and Filtering.** Concepts selected by the above approach need to be post-processed.

Many of the action concepts given by UCF 101 are not directly related to the events we want to classify. However, these actions capture some basic motion characteristics and were selected as discriminative concepts. For example, *high jump* was often selected for *winning a race without a vehicle* event, most likely due to it involves person running outdoors. We mapped the action concept names of UCF 101 to more general action names, some examples are shown in Table 2.

The mapping can also be used for motion and visual concepts other than UCF 101, and be event specific. For example, *military aircraft* was mapped to *vehicle* for *changing tire* event, as it was often selected for video segments with *vehicles*.

Even after name mapping, there were still some concepts being selected that did not appear in the video segments (e.g. *eukaryotic organism*) due to the concept detector noise. We used an event specific white list to filter the selected concepts: we first retrieved a list of concepts with high occurrences for an event. It was done by setting up a validation dataset independent from testing dataset. We then manually selected the concepts which are related to the event, and added them to the white list. During testing, concepts that were selected but not in the white list were dropped.

Finally, we used a simple heuristic to select at most two object concepts, one action concept and one scene concept.

Table 3 lists the visual and motion concepts with the most occurrences after ranking, mapping and filtering.

| Event name | Concept 1 | Concept 2 | Concept 3 |
|---|---|---|---|
| Birthday party | Blowing candles | Male | Clapping |
| Changing tire | Turning wrench | Using tube | Outdoor |
| Flash mob | Dancing | Crowd | Marching |
| Vehicle unstuck | Vehicle moving | Vehicle | Outdoor |
| Grooming animal | Washing | Hands | Animal |
| Making sandwich | Spreading | Food | Kitchen |
| Parade | Marching | Streets | Crowd |
| Parkour | Running | Climbing | Building |
| Repairing appliance | Amateur video | Repairing | Hands |
| Sewing | Sewing | Person | Hands |
| Bike trick | Horizontal move | Person | Outdoor |
| Cleaning appliance | Appliance | Washing | Wiping |
| Dog show | Walking | Person | Field |
| Giving directions | Person | Walking | Suburban |
| Marriage proposal | Hugging | Kissing | Person |
| Renovating home | Horizontal move | Using tool | Indoor |
| Rock climbing | Rock climbing | Trees | Outdoor |
| Town hall meeting | Talking | Crowd | Flags |
| Winning race | Horizontal move | Sports | Racing |
| Metal crafts project | Man-made thing | Hand move | Glow |

**Table 3: Visual and motion concepts selected the most times by our visual observation generation system on MEDTest dataset**

### 4.1.2 ASR Observations

For generating MER observations based on semantic concepts from ASR, we used a cascading approach. Each video was segmented into windows of 1 second duration for localizing MER observations. Lattice n-gram counts were collected for the entire video as well as for each window. The ASR MED system was based on a sparse linear kernel SVM model with an $l$-1 penalty. The features from the SVM model are log-mapped unigram lattice n-gram counts from ASR. From the SVM model, we can obtain the most significant features, i.e., functions of unigram counts $C(w)$, with their weights $\alpha_E(w)$ for each event $E$, where and $W_E$ is the set of important words for $E$. Considering how the features for SVM training were generated, the importance score $\theta_E(w)$ for an important word is computed as

$$\theta_E(w) = \alpha_E(w) \cdot \log(1 + 10000C(w)) \tag{3}$$

For each trial of video and event pair as $< V, E >$, we generated the ASR lattice n-gram counts for $V$ and identified the subset of $W_E$ appearing in the counts. We denote this subset as MER words for trial $< V, E >$. For each of these MER words, we identified the 1-second window(s) where this word appears and selected the window with the highest ASR confidence score for this word as the MER snippet for the word. If there are multiple MER words in one 1-s interval, they are presented by ranking based on the ASR confidence scores, from high to low. The importance score $\theta_E(w)$ is presented as the importance value and the ASR confidence score is presented as the confidence value. We denote MER results from this first step as MER candidates.

We then ran a set of post-processing filtering modules as the second step on the MER candidates to generate more semantically oriented and coherent MER observations. We used a Wordnet based filtering module for 2013 MER evaluation and developed a topic-modeling based filtering module after evaluation. The Wordnet based filtering module applies Porter stemming on the MER candidates, interfaces with Wordnet to identify whether a stemmed MER candidate is noun or verb, and then further filters out auxiliary verbs. In post-evaluation development, we developed a topic modeling based filtering algorithm for ASR MER presentation. We experimented with learning topics combining event kit text and learned significant ASR concepts from the ASR MED system. We compared various topic modeling approaches such as LDA and cross-entropy and cosine similarity measures. We used the topic clusters to identify words that are semantically distant from the cluster centroids (keywords). Preliminary experiments showed that this approach is effective to identify important ASR concepts for MED that were introduced due to homophones. For example, *chis* in the ASR vocabulary is a homophone for *cheese*, and due to noisy ASR training transcriptions, it appears frequently in recognition output for videos of *making a sandwich*. This approach could identify words like *chis* as outliers and replace them with their homophone candidates *cheese* which are much closer to the cluster centroid.

### 4.1.3 OCR Observations

After text recognition, the resulting words were aggregated into coherent phrases and the importance of the phrases relative to the event class were then scored. High scoring phrases were considered observations. Because each line of video text was recognized independently, independent detections were grouped together into a single phrase if the amount of time between the two pieces of recognized text was less than 30 ms. Thus multiple lines of text that appeared either in a single frame, or in a contiguous succession of frames, were treated as a single coherent group and collectively used to determine their importance. The importance of a phrase was defined as the posterior probability of the event class given the phrase words (removing stopwords), estimated using logistic regression.

### 4.1.4 Selecting and Merging Observations

MER observations were generated from three sources: visual/motion concepts, ASR and video OCR. Each source also scored the importance of their observations. One possible strategy to merge the observations is to learn the mapping from estimated importance scores to actual human importance, normalize the scores across the three sources, and then select the most importance subset. Unfortunately, given our training videos, there were not enough observations generated by ASR and OCR in all the event classes to properly model the importance. Thus, we decided to select observations from each source independently.

There were enough visual and action concept observations to determine their usefulness. We generated observations for 20 videos per event, where the videos were drawn from a validation set. We then had one human subject annotate each observation interval as complete (in determining the event class), relevant (useful but not complete) or not useful. It was determined that selecting the top four visual/action concept observations per video generated a complete set of evidence a high percentage of the time, so this strategy was used.

By hand, we determined subjectively that same strategy did not work for ASR and OCR; there were too many irrelevant observations generated by them. For ASR this was dealt with by adding an importance score threshold, hand-selected for each event, in addition to a count limit of four. For several of the event classes, the importance was considered too unreliable and the threshold was effectively set to infinity.

For OCR, the importance scores were too unreliable for too many of the event classes, so keyword matching was substituted for importance thresholding. For each event, a keyword profile was generated from the provided text description. The relevance of each word was estimated using TF-IDF, the words were sorted by relevance and all the words above a hand-selected relevance threshold were used as the keyword profile for that event. An OCR observation was selected if a substring of its associated text phrase matched any keyword in the profile. At most, four OCR observations were selected (the top four importance scores after keyword matching).

Once the observations from each source were selected, they were ordered chronologically in the MER document.

Table 4 shows the number of selected observations from different modalities. It is clear that most videos had visual observations selected. There were a reasonable amount of videos with ASR and OCR detections, but only a small proportion were selected in MER observations.

| Event | Making a sandwich | Parkour |
|---|---|---|
| # True positives | 140 | 104 |
| # SESAME positives | 314 | 244 |
| Videos with visual MER | 305 | 238 |
| Videos with any ASR | 274 | 177 |
| Videos with ASR in MER | 147 | 0 |
| Videos with any OCR | 206 | 139 |
| Videos with OCR in MER | 27 | 12 |

**Table 4: MER statistics for two MEDTest events. Most of the detected positives have visual observations. Only a proportion of ASR and OCR detections were used in MER.**

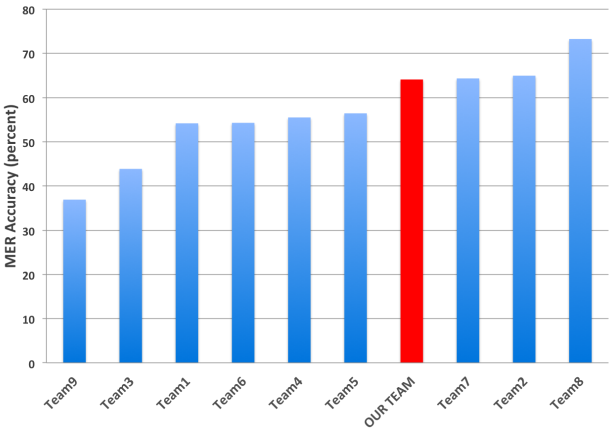


**Figure 3: Accuracy comparison (The higher the better) on TRECVID 2013 MER task.**

## 4.2 Performance Evaluation

The purpose of MER is to help users accurately and rapidly assess whether each clip is truly a positive for a specified event. To assess progress towards these objectives, NIST [13] developed three metrics and manually judged each TRECVID MER submission in terms of them. The judges decided whether a clip was a positive for an event using only the system-generated MER observations, where each observation had a selected clip segment and an associated text description. Figure 6 gives four sample MER results generated by ISOMER: The top three results provided reliable observations from visual concepts, ASR and OCR, respectively. The bottom one is a failure case where wrong concept names were returned.

The metrics were the following:

**Accuracy**: the percent of correctly judged clips (agreeing with the MED ground truth).

**Percent Recounting Review Time (PRRT)**: the percentage of clip time the judges took to perform the assessment. (100% means that the judges, using only MER for information, took as long as the video itself.)

**Precision of the observation text**: a subjective measure of how well the observations described the segments. The judges scored each observation as excellent (4), good

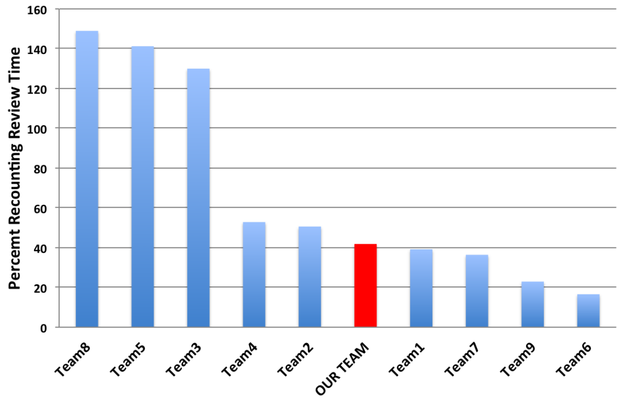


**Figure 4: Precision Recounting Time comparison (The lower the better) on TRECVID 2013 MER task.**

(3), fair (2), poor (1), or fails (0); precision was the mean score for a MER submission.

Clearly there is a tradeoff between accuracy and review time: the ideal judge, watching the whole video, should be perfectly accurate (100%). To encourage the developers to produce a concise set of observations optimized for both accuracy and speed, NIST had the judges review every observation generated for a clip before making a decision.

Ten teams submitted MER for TRECVID 2013 [1], their scores for each metric are shown in Figure 3 to Figure 5. Our system achieved an accuracy of 64.1%, a precision of 2.53 (out of a possible 4.0), and a Percent Recounting Review Time of 41.8%. Of the ten submissions, our system produced the highest precision. It was the only one with *good* (3.0) as the closest qualitative category to its average; the others ranged from *poor* (1.0) to *fair* (2.0). We were also competitive in the accuracy-speed tradeoff. The system with the highest accuracy (73.26%) took 149% of the time to review (49% longer to evaluate than just watching the whole video). The other top accuracy scores (64.96%, 64.34%) were essentially the same as ours (64.1%) and had similar review times (50.6% and 36.4%) to our 41.8% of the video.

Although many of the visual concepts cited in the MER observations were not present in the associated video snippet, merely choosing the most relevant interval for viewing is often very helpful to the user. Using another annotator, we performed an additional manual examination of 250 observations, and found that 75% of the associated three-second video segments contained enough content for a user to determine whether the video label is correct. This is in rough agreement with our other assessment discussed earlier. Overall, our system's performance points to the merits of the ISOMER approach where we first select the best interval and then determine the best semantic concepts within the interval to display.

## 5. ACKNOWLEDGEMENT

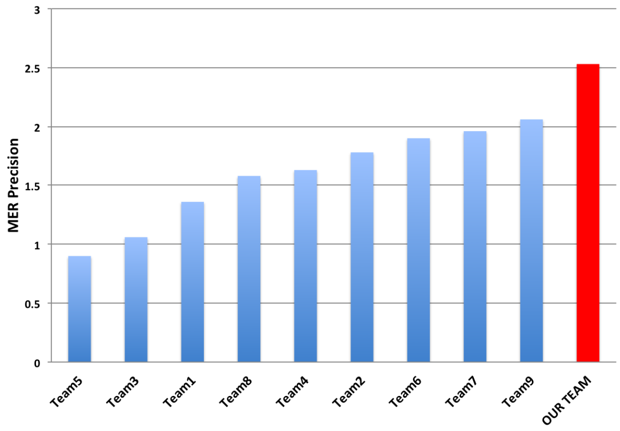


**Figure 5: Text precision comparison (The higher the better, best is 4.0) on TRECVID 2013 MER task.**

This work was supported by the Intelligence Advanced Research Projects Activity (IARPA) via the Department of Interior National Business Center (DoI/NBC), contract number D11PC0067. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, or the U.S. Government. Computation for the work described in this paper was supported, in part, by the University of Southern California Center for High-Performance Computing and Communications (`hpcc.usc.edu`).

## 6. REFERENCES

[1] M.-Y. Chen and A. Hauptmann. Mosift: Recognizing human actions in surveillance videos. Technical Report CMU-CS-09-161, 2009.

[2] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR*, 2009.

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

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

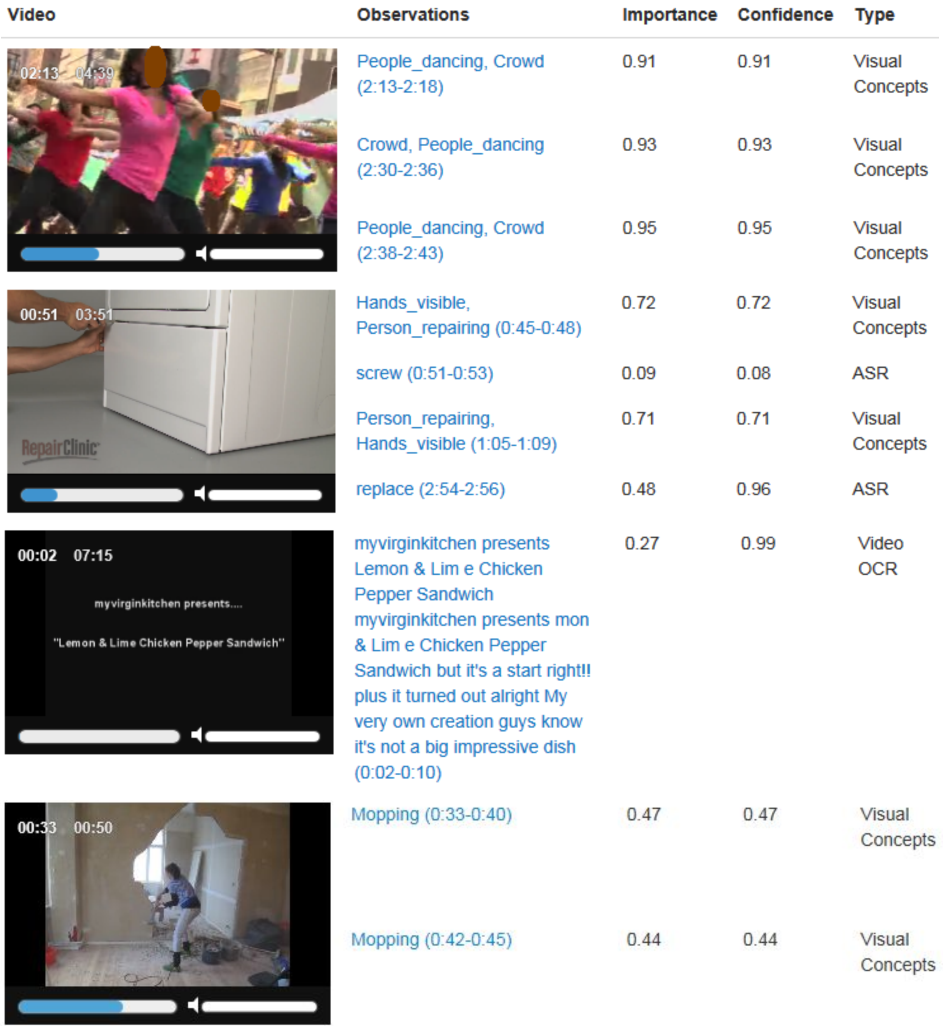
[5] H. Izadinia and M. Shah. Recognizing complex events using large margin joint low-level event model. In *ECCV*, 2012.

[6] H. Jégou, M. Douze, and C. Schmid. Product quantization for nearest neighbor search. *TPAMI*, 2011.

[7] I. Laptev. On space-time interest points. *IJCV*, 2005.

[8] 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] Results and system descriptions for other participants can be found at `http://www-nlpir.nist.gov/projects/tvpubs/tv.pubs.13.org.html`



02:13 04:39
00:51 03:51
RepairClinic
00:02 07:15
myvirginkitchen presents....
"Lemon & Lime Chicken Pepper Sandwich"
00:33 00:50

| Video | Observations | Importance | Confidence | Type |
|---|---|---|---|---|
| | People_dancing, Crowd (2:13-2:18) | 0.91 | 0.91 | Visual Concepts |
| | Crowd, People_dancing (2:30-2:36) | 0.93 | 0.93 | Visual Concepts |
| | People_dancing, Crowd (2:38-2:43) | 0.95 | 0.95 | Visual Concepts |
| | Hands_visible, Person_repairing (0:45-0:48) | 0.72 | 0.72 | Visual Concepts |
| | screw (0:51-0:53) | 0.09 | 0.08 | ASR |
| | Person_repairing, Hands_visible (1:05-1:09) | 0.71 | 0.71 | Visual Concepts |
| | replace (2:54-2:56) | 0.48 | 0.96 | ASR |
| | myvirginkitchen presents Lemon & Lim e Chicken Pepper Sandwich myvirginkitchen presents mon & Lim e Chicken Pepper Sandwich but it's a start right!! plus it turned out alright My very own creation guys know it's not a big impressive dish (0:02-0:10) | 0.27 | 0.99 | Video OCR |
| | Mopping (0:33-0:40) | 0.47 | 0.47 | Visual Concepts |
| | Mopping (0:42-0:45) | 0.44 | 0.44 | Visual Concepts |

**Figure 6: More results (*flash mob*, *repairing an appliance*, *making a sandwich* and *repairing an appliance*) from our ISOMER system.**

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

[10] G. K. Myers, R. C. Bolles, Q.-T. Luong, J. A. Herson, and H. Aradhye. Rectification and recognition of text in 3-d scenes. *IJDAR*, 2005.

[11] G. K. Myers, R. Nallapati, J. van Hout, S. Pancoast, R. Nevatia, C. Sun, A. Habibian, D. C. Koelma, K. E. A. van de Sande, A. W. M. Smeulders, and C. G. M. Snoek. Evaluating multimedia features and fusion for example-based event detection. *MVA*, 2014.

[12] P. Natarajan et al. Bbnviser : Bbn viser trecvid 2012 multimedia event detection and multimedia event recounting systems. In *TRECVID*, 2012.

[13] P. Over, G. Awad, M. Michel, J. Fiscus, G. Sanders, W. Kraaij, A. F. Smeaton, and G. Queenot. Trecvid 2013 – an overview of the goals, tasks, data, evaluation mechanisms and metrics. In *TRECVID*, 2013.

[14] J. Sanchez, F. Perronnin, T. Mensink, and J. Verbeek. Image Classification with the Fisher Vector: Theory and Practice. *IJCV*, 2013.

[15] A. F. Smeaton, P. Over, and W. Kraaij. Evaluation campaigns and trecvid. In *MIR*, 2006.

[16] 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*.

[17] C. Sun and R. Nevatia. Active: Activity concept transitions in video event classification. In *ICCV*, 2013.

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

[19] B. T. Truong and S. Venkatesh. Video abstraction: A systematic review and classification. *TOMCCAP*, 2007.

[20] K. E. A. van de Sande, T. Gevers, and C. G. M. Snoek. Evaluating color descriptors for object and scene recognition. *TPAMI*, 2010.

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