

BBN VISER TRECVID 2013 Multimedia Event Detection and Multimedia Event Recounting Systems

Pradeep Natarajan, Shuang Wu, Florian Luisier, Xiaodan Zhuang, Manasvi Tickoo
Speech, Language, and Multimedia Business Unit, Raytheon BBN Technologies, Cambridge, MA, USA

Guangnan Ye, Dong Liu, Shih-Fu Chang
*Department of Electrical Engineering
Columbia University, New York, USA*

Imran Saleemi, Mubarak Shah
*Department of Electrical Engineering and Computer Science
University of Central Florida, Orlando, USA*

Vlad Morariu, Larry Davis
*Department of Computer Science
University of Maryland, USA*

Abhinav Gupta
*Robotics Institute,
Carnegie Mellon University, USA*

Ismail Haritaoglu
*Polar Rain Inc.
Menlo Park, USA*

Sadiye Guler, Ashutosh Morde
*IntuVision Inc.
Woburn, MA, USA*

Abstract

We describe the Raytheon BBN Technologies (BBN) led VISER system for the TRECVID 2013 Multimedia Event Detection (MED) and Recounting (MER) tasks. We present a comprehensive analysis of the different modules: (1) a large suite of visual, audio and multimodal low-level features; (2) video- and segment-level semantic scene/action/object concepts; (3) automatic speech recognition (ASR); (4) videotext detection and recognition (OCR). For the low-level features, we used multiple static, motion-based, color, and audio features and Fisher Vector (FV) representation. For the semantic concepts, we developed various visual concept sets in addition to multiple existing visual concept banks. In particular, we used BBN's natural language processing (NLP) technologies to automatically identify and train salient concepts from short textual descriptions of research set videos. We also exploited online data resources to augment the concept banks. For the speech and videotext content, we leveraged rich confidence-weighted keywords and phrases obtained from the ASR and OCR systems. We combined these different streams using multiple early (feature-level) and late (score-level) fusion strategies. Our system involves both SVM-based and query-based detections, to achieve superior performance despite of the varying number of positive videos in the event kit. We present a thorough study of different semantic feature based systems compared to low-level feature based systems.

Consistent with previous MED evaluations, low-level features still exhibit strong performance. Further, our semantic feature based systems have improved significantly, and produce gains in fusion, especially in the EK10 and EK0 conditions. On the pre-specified condition, the mean average precision (MAP) of our VISER system are 33%, 16.6% and 5.2% for the EK100, EK10 and EK0 conditions respectively. These are largely consistent with our ad hoc results that are 32.2%, 14.3% and 8.1% for the EK100, EK10 and EK0 conditions respectively. For the MER task, our system has an accuracy of 64.96% and takes only 52.83% of the video length for the evaluators to analyze the evidence and make their judgment.

Description of Submitted Runs

Pre-Specified Task:

EK100:

BBNVISER_MED13_OCRSys_PROGAll_PS_100Ex_1:

This system combines OCR systems trained on the provided EK100 examples and query based systems that utilize the provided event kit text.

BBNVISER_MED13_ASRSys_PROGAll_PS_100Ex_2:

Similar to the OCR submission, this submission combines an ASR system trained with the EK100 examples with a query based system.

BBNVISER_MED13_VisualSys_PROGAll_PS_100Ex_1:

This system combines multiple sub-systems developed using low-level and semantic visual features that are trained using the EK100 examples.

BBNVISER_MED13_AudioSys_PROGAll_PS_100Ex_1:

This system combines multiple sub-systems developed using low-level audio features that are trained using the EK100 examples.

BBNVISER_MED13_FullSys_PROGAll_PS_100Ex_1:

This system combines multiple sub-systems based on low and high-level visual and audio features, as well as ASR and OCR based systems.

EK10:

The BBNVISER_MED13_OCRSys_PROGAll_PS_10Ex_1, BBNVISER_MED13_ASRSys_PROGAll_PS_10Ex_2,

BBNVISER_MED13_VisualSys_PROGAll_PS_10Ex_1, BBNVISER_MED13_AudioSys_PROGAll_PS_10Ex_1,

BBNVISER_MED13_FullSys_PROGAll_PS_10Ex_2 are similar to the EK100 counterparts but are trained using the provided

EK10 examples.

EK0:

The BBNVISER_MED13_OCRSys_PROGAll_PS_0Ex_1, BBNVISER_MED13_ASRSys_PROGAll_PS_0Ex_1, BBNVISER_MED13_VisualSys_PROGAll_PS_0Ex_1, BBNVISER_MED13_AudioSys_PROGAll_PS_0Ex_1, BBNVISER_MED13_FullSys_PROGAll_PS_0Ex_1 are based on the OCR, ASR and semantic audio and visual features, and utilize a query based approach for event detection.

Ad Hoc Task:

We submitted a set of systems identical to the pre-specified task for the ad hoc task.

1 Introduction

Techniques for fast, automatic analysis of large volumes of unconstrained web videos and detection of events of interest has several compelling applications. The annual TRECVID Multimedia Event Detection (MED) and Recounting (MER) evaluations [Smeaton et al. 2006, Over et al. 2013] aim to measure progress in developing such techniques, and strong performance has been reported in recent evaluations [Jiang et al. 2010, Bao et al. 2011, Natarajan et al. 2011, Aly et al. 2012]. The core of these systems is based on the *bag-of-words* [Csurka et al. 2004] approach built on low-level features extracted from pixel patterns in videos. This approach has several advantages, such as compact video representation and ease of model training and prediction using well understood techniques such as support vector machines (SVM). However, this method requires a large training set to train reliable event detectors. Furthermore, this approach does not provide the ability to recount or reason about the events and evidences seen in videos.

In this paper, we provide an overview of the BBN's VISER system for the TRECVID 2013 MED and MER evaluations. Our system uses a combination of low-level visual, audio and multimodal features, as well as semantic audio-visual concept detectors. Our low-level system combines a set of features that capture salient gradient, color, texture, motion and audio patterns. Our semantic system includes a suite of off-the-shelf detectors as well as a novel set of weakly supervised concept detectors. We also leverage data downloaded from the web for training our concept detectors. For the MED task, we fused these with speech and videotext output using late fusion to get final system outputs. For the MER task, the outputs of the semantic concept detectors along with speech and videotext were thresholded and combined to produce event recounting. Our main findings can be summarized as follows:

- Low-level features continue to exhibit strong performance and form the core of our EK100 and EK10 submissions;
- Semantic features produce significant gains for the EK10 system and also show promising performance for the EK0 condition;
- Speech and videotext provide complementary evidences and consistently improve performance for both pre-specified and ad hoc event detection;
- Our MER system allows an analyst to achieve reasonable event detection accuracy and the evidences take 50% less time to view than the full video.

The rest of the paper is organized as follows. In Section 2, we describe our low-level feature system in detail. Section 3 describes our high-level semantic features system. In Sections 4 and 5, we describe our ASR and videotext OCR systems. In Section 6, we present the different feature fusion strategies. We conclude with a discussion of experimental results in Section 7.

2 Low-level Features

We extracted and combined multiple audio and visual features using Fisher Vector encoding. We considered the following features and systems:

Audio Features: We considered Mel-Frequency Cepstral Coefficient (MFCC), Frequency Domain Linear Prediction (FDLP) and Audio Transients (AT) features. In addition, we tested a low-level audio feature (LLfeat-A) system.

Visual Features: We considered multiple visual features including a Compressed Histogram of Oriented Gradients (CHOG) [Chandrasekhar et al. 2011], gray-scale dense-SIFT (D-SIFT) [Boureau et al. 2010], color D-SIFT (Opp-D-SIFT) [van de Sande et al. 2010] computed in the opponent color space, BBN's kernel descriptor-based motion feature (KDES-GF1) [Natarajan et al. 2012], a dense trajectory-based feature that combines shape and motion features (DT) [Wang et al. 2013]. In addition, we developed a system that combines all these low-level visual features (LLfeat-V).

Audio-Visual Fusion: We also combined the audio and visual features to get a single system called LLfeat.

3 Semantic Features

Ability to detect high-level semantic concepts in videos is crucial for event recounting and event detection with a small training set. However, there are several challenges in developing robust systems for detecting such semantic concepts. First, the set of possible concepts that can occur in web videos is potentially infinite, making traditional ontology based approaches such as LSCOM infeasible. Second, it is extremely time consuming to collect a sufficient number of annotations for concepts of interest

that can be used to train robust concept detectors. The annotation task becomes harder if it involves marking spatial or temporal bounding boxes.

3.1 Evaluation of Off-the-Shelf Concept Detectors

We conducted a detailed evaluation of two popular off-the-shelf concept detectors for the MED task: ObjectBank [Li et al. 2010] and SUN scene attribute [Patterson et al. 2012] features. SUN scene features produce a feature vector where each element corresponds to the detection score for a particular concept. ObjectBank uses a spatial pyramid representation and produces detection confidence scores at different spatial patches for each concept. In this case, we measured the performance of both the full feature vector (Full), as well as the vector obtained by considering only the highest confidence detection for each concept (Max). The table below summarizes the performance of these features compared to our baseline dense SIFT based low-level feature.

Features	AP	R0
D-SIFT (baseline)	0.2779	0.3776
SUN Scene Attributes	0.1041	0.1384
ObjectBank (Full)	0.1565	0.1893
ObjectBank (Max)	0.0459	0.0508

Table 1: Comparison of ObjectBank and SUN Scene attribute features for MED for EK100 training condition on BBN’s internal test partition.

The off-the-shelf detectors we tested had significantly weaker performance compared to low-level features for the EK100 training conditions, and also did not help in fusion.

3.2 Weakly Supervised Concepts (WSC)

We developed techniques to exploit the annotations in the judgment files provided by LDC for training concept detectors. This allows us to utilize annotations of the research set videos that are already available at no additional cost. However, a challenge with this data is that they are short, free-form text descriptions. We address this by applying BBN’s natural language processing technologies to detect salient concepts in the text annotations.

For each of these concepts, we aggregated the corresponding videos in whose annotations they occurred. Then, we pruned all concepts that had too few video occurrences to ensure we had sufficient examples to validate and train the concept detectors. Next, we extracted multiple low-level features from the videos to capture salient gradient, motion, color and audio patterns. We then trained detectors for each concept by combining these features. Finally, we did a second round of pruning the concepts based on the mean average precision (MAP) metric to ensure that the detected concepts have a reasonable level of accuracy. The detected concepts can be directly used for recounting (MER) and describing the video. Further, for the event detection task (MED), we use the vector of detection scores for different concepts for training event detectors. The table below illustrates the performance of these features on the official MEDTEST partition:

Features	AP	R0
D-SIFT	0.2897	0.3838
WSC-D-SIFT	0.2108	0.3405
DT	0.3357	0.4272
WSC-DT	0.2718	0.3817

Table 2: Comparison of weakly supervised concepts with low-level features on MEDTEST EK100. WSC-D-SIFT and WSC-DT refer to WSCs trained on D-SIFT and the dense trajectory (DT) features respectively.

As can be seen, WSCs are significantly weaker than the corresponding low-level feature based system, but are stronger than off-the-shelf concept detectors. Further, we tested combining these concept detectors with a low-level feature based system. We found that they produce strong performance gains for the EK10 condition, while the gains for EK100 are modest. This is illustrated in the table below:

Features	AP	R0
EK10-LLfeat-V	0.1459	0.1885
EK10-LLfeat-V + WSC-D-SIFT + WSC-DT	0.1785	0.2190
EK100-LLfeat-V	0.3810	0.4771
EK10-LLfeat-V + WSC-D-SIFT + WSC-DT	0.3852	0.4830

Table 3: Fusion of WSC features with a low-level feature based visual system (LLfeat-V) for the EK10 and EK100 conditions.

3.3 Concept Discovery from Flickr

We have designed a novel framework for automatic concept discovery from the internet images.

(1) **Candidate Concept Discovery:** We first retrieve a set of images from Flickr, and re-rank them by the event SVM model trained with the TRECVID MED training videos so that the top ranked images are visually related to the target event videos. We calculate TF-IDF values of the tags associated with these top ranked images and treat the tags with high TF-IDF values as the candidate concept pool.

(2) **Visual Concept Verification:** To ensure that only visual related concepts are included, we first treat the images associated with a concept as positive training samples and choose a number of images from the other concepts as negative training samples. Then we split all training images into two halves and do 2-fold cross validation. Finally, only the concepts with high cross validation performance are verified as visually related concepts and retained in the concept library.

(3) **Concept Representation Generation:** We consider each tag as a concept and use the associated images to train a SVM concept classifier. The concept scores generated on the MED videos are then treated as concept based video representation.

(4) **Salient Concept Selection based on L_1 -SVM:** After the concept representation generation, each video is represented as a concept score vector. However, some dimensions may still be irrelevant to the target event. To remove such noisy dimensions, we use L_1 -SVM to automatically set the values of irrelevant concept dimensions as zero. Once we select the salient concepts from L_1 -SVM, we can represent all videos based on this concept subset and train an event detection model using L_2 -SVM.

Our method has several advantages:

1. Concepts are automatically discovered from the internet, and most of them are highly relevant.
2. Images for training a concept classifier are chosen to be more compatible with video contents, and thus the content discrepancy between different visual resources is reduced.
3. L_1 -SVM allows us to optimally determine the concept descriptions from the entire concept pool, and achieves robust semantic based event detection.
4. Our concept classifiers can also be used for the MER task. Specifically, for each video, we ranked the concept scores and treated the top ranked concepts as the recounting concepts of the video content. Since the concepts classifiers are trained separately, the SVM scores cannot be compared directly. We thus applied Gaussian normalization to make the scores comparable.

3.4 Object Detection

We used detections from a state-of-the-art object detector developed by Pedro Felzenszwalb at the University of Chicago [Felzenszwalb et al. 2010]. We used a representation called the *spatial probability map* which captures the spatial distribution of an object’s presence in a video. Overall, we found car detections to produce consistent gains for the “*Getting vehicle unstuck*” event, but did not find significant improvement when we used other detectors. The person detections provided salient information for the recounting task.

3.5 Salient Object-based Concept Feature

We also applied the Classemes models provided in [Torresani et al. 2010] to generate novel scene concept features. These models were trained over a large scale concept pool (around 3,000 concepts) defined in LSCOM. In order to refine the concept feature output, we proposed the idea of a salient object based concept feature. Specifically, we first detect regions containing prospective salient objects based on image multi-scale saliency, color contrast, edge density and straddleness. Within each region, we use the Classeme concept detector. Max-pooling is used for each frame result. Average-pooling is used for multiple frames within each video. This approach consistently improved performance over the Classeme baseline in our experiments.

4 Automatic Speech Recognition

We use GMM-based speech activity detection (SAD) and HMM-based multi-pass large vocabulary automatic speech recognition (ASR) to obtain speech content in the video, and encode the hypotheses in the form of word lattices. We first transform the raw audio into a 45 dimensional feature stream using the following steps. 14 Mel-warped cepstral coefficients were extracted from overlapping frames of audio data, each 29ms long, at a rate of 100 frames per second. Each segment of speech is normalized by the mean cepstrum and peak energy non-causally, removing any long term bias due to the channel. In addition, the feature vectors are scaled and translated such that for each video, the data has zero mean and unit variance. These 14 base cepstral features and the energy, together with their first and second derivatives, compose the final 45-dimensional feature vector.

Then, the speech segments are identified by the SAD system [Ng et al. 2012]. The SAD system employs two Gaussian mixture models (GMM), for speech and non-speech observations respectively. A small subset of 100 video clips was annotated for speech segments, used for training the speech GMM. Besides the non-speech segments in this set, we also use 500 video clips with music content to enrich the non-speech model, in order to handle the heterogeneous audio in consumer video. SAD was evaluated on 40 video clips and obtained a False Alarm rate of 10.1% and Miss Detection of 5.8% according to the NIST *md-eval* metric, with a 0.25 seconds collar.

Given the automatically detected speech segments, we then apply BBN's large-vocabulary ASR system to the speech data to produce a transcript of the spoken content. This system is adapted from an ASR system trained on 1,700-hour English Broadcast News. In particular, we update the lexicon and language model using MED 2011 descriptor files [Over et al. 2011], relative web text data, and the small set of 100 video clips with annotated speech transcription. We use a trigram language model trained over 2 million words of in-domain data from the MED 2011 descriptor files and relative web text data and 11 billion words of out-of-domain web-data. The vocabulary size is about 168k. The acoustic models are adapted during ASR decoding for each video clip in an unsupervised fashion via Maximum Likelihood Linear Regression (MLLR) and Constrained MLLR (CMLLR). We evaluated the baseline ASR model and adapted ASR model on a held-out set of 100 video clips from the MED 2011 set [Over et al. 2011]. The WER of the baseline system was 48.2% and the WER of the adapted system was 35.8%. The system outputs not only the 1-best transcripts but also word lattices with acoustic and language model scores.

5 Videotext OCR

A videotext detector detects the bounding boxes, whose content is recognized by HMM-based multi-pass large vocabulary OCR. Similar to the ASR system, word lattices are used to encode alternative hypotheses. We leverage a statistically trained videotext detector based on SVM to estimate videotext bounding boxes. This detector is developed based on [Peng et al. 2011] and is briefly summarized here.

Maximally Stable Extremal Regions (MSER) which are robust to illumination and viewpoint variations are selected as text candidates. Rich shape descriptors such as Histogram of Oriented Gradients (HOG), Gabor filters, corners and geometrical features are used to represent the candidates and classified using a support vector machine (SVM). Each positively labeled candidate serves as anchor region for word formation. We then group candidate regions based on geometric and color properties to form word boundaries. This allows us to overcome the mistakes of the classification step. To speed up the system for practical applications while preserving discriminative features, we use Partial Least Squares (PLS) approach for dimensionality reduction. The detected words are binarized and filtered before being passed to an HMM-based OCR system for recognition. On a small consumer video dataset with videotext bounding boxes annotated, the videotext detector achieves pixel-level precision and recall of 67.9% and 31.8%. Note that these measurements are calculated on the raw pixel level, as our HMM-based OCR system expects tight bounding boxes around videotext regions.

With each identified videotext bounding box, this two-dimensional text image is converted to a one-dimensional frame sequence, each frame characterized by features such as intensity percentile, local stroke angle, correlation and total energy within the sliding window corresponding to the frame. Then the HMM-based BBN OCR system finds a sequence of characters that maximizes the posterior, by using glyph models (similar to the acoustic models in ASR), a dictionary and N-gram language models. This OCR system employs various parameter sharing and performs recognition in a multi-pass fashion. The vocabulary size is about 140k. Details about the BBN HMM-based OCR systems can be found in [Peng et al. 2013]. Since the videotext content presents itself in various forms, such as subtitles, markup titles, text in scenes (e.g., banners and road signs), it is much more challenging than conventional scanned document OCR. Considering that we focus on bag-of-words representation for OCR in this work, we measure the word precision and recall within each video, at 37% recall and 14.7% precision.

6 Classifier Learning and Feature Fusion

Using the features described so far, we built multiple sub-systems by training kernel based classifiers for each event. During this process, we jointly optimized the classifier parameters and the detection threshold. Given a test video, we obtained classification scores for each of these sub-systems. We then applied a late fusion strategy to combine these scores and obtain a final detection score. During training, we also estimated a detection threshold for the late fusion system. In this section, we will describe each of these steps.

6.1 Early Fusion

We trained different subsystems by combining different features from the same class, such as appearance, color, motion, etc. For our EK100 and EK10 systems, we first computed χ^2 kernels for each feature and then combined them using kernel product. Further, we used standard parameter estimation techniques to optimize the performance of each sub-system.

6.2 System Combination

After training the different sub-systems and estimating their detection thresholds, we combined the different sub-systems using weighted average fusion. Here, in addition to computing a global system level weight, we adaptively weight each system's output on a video by video basis. The first is a system level weight (w_l), which was calculated from the ANDC scores of each system

based on our internal partitions. The second is a video specific weight (w_2), calculated from the optimal threshold for the system found during our threshold analysis, and the confidence score for a given test video.

Given these weights, the output score P for a video j is simply given by:

$$P(j) = \frac{\sum_i w_1(i)w_2(i,j)p_{ij}}{\sum_i w_1(i)w_2(i,j)} \quad (3)$$

We conducted a series of experiments to identify the optimal combination of systems for the EK100, EK10 and EK0 conditions. For each training condition, we obtained multiple sub-systems by combining individual features and then combined the sub-systems using the weighted average based fusion strategy.

Late fusion of multiple sub-systems consistently improves performance of our EK100 and EK10 systems compared to the best single sub-system. Further, fusion of EK0 systems with EK10 produced additional gains. Based on the combinations we identified, we submitted our systems to NIST for the TRECVID MED pre-specified and ad hoc conditions.

6.3 0-shot Retrieval

For our 0-shot system, we compared multiple audio and visual semantic feature based systems. For the audio modality, we considered outputs from automatic speech recognition (ASR), as well as the WSC features previously described, trained with the audio MFCC features. For the visual modality, we combined outputs from WSC-D-SIFT, WSC-DT, ObjectBank, SUN Scene Attributes, Classemes, as well as features learned from images downloaded from YouTube, Google Images and Flickr. We also leverage video text output from two different systems: one that performs character-level decoding (and hence can produce spelling errors but can capture useful content in non-standard text such as website names), and one that performs word-level decoding (that avoid spelling errors but produces erroneous outputs with non-standard text).

In addition, we also developed a novel approach for leveraging low-level features for the 0-shot problem. We built this by utilizing the video-level text annotations available for the research set. After identifying salient concepts in the text, we averaged the low-level feature vector of all the videos in which each concept occurs. At query time, we retrieve the average vectors of the concepts in the query, compute the average of these vectors and then rank the search collection based on distances to this average vector.

As expected, ASR and OCR have the strongest individual performances. We then combined these features for the official ASR, Audio, OCR, Visual and Full system submissions. An important challenge in 0-shot is threshold estimation, since we do not have any positive examples to calibrate the threshold. We addressed this challenge by setting the threshold based on a fixed false alarm rate computed on the background set for different systems.

7 Experiments and Results

In this section, we present results for the different systems we submitted for the Pre-Specified and Ad Hoc MED tasks, and for MER, on the PROGTEST set.

7.1 Pre-Specified Event Detection Submission Systems

For the pre-specified task, we submitted systems for all the training conditions and contrastive runs for ASR, OCR, Audio-only, and Visual-only systems besides the Full systems. The table below presents the MAP scores for the different submissions.

	FullSys	ASRSys	AudioSys	OCRSys	VisualSys
EK100	33.0%	7.6%	12.0%	4.8%	28.2%
EK10	16.6%	3.5%	4.4%	3.2%	13.3%
EK0	5.2%	1.4%	0.5%	2.8%	3.5%

Table 4: MED 2013 Pre-Specified Results

As expected, there is a large drop in performance from EK100 to EK10 and EK0. However, even for EK0 the performance of the full system performs significantly better than random. The visual-only sub-system is the strongest individual system even for the EK0 condition, illustrating the utility of our semantic features. Combination of the visual system with the other systems consistently produces performance gains.

Our pre-specified submissions overall had strong performance in comparison to other TRECVID 2013 submissions. Our system had the top performance among all submissions for EK10-FullSys, 2nd for EK100-FullSys and 3rd for EK0-FullSys. We also had top performance for EK10 and EK10 Visual Systems and EK100 OCR system. Our systems were in the top-2 for most of the remaining contrastive runs and top-3 in all the submissions.

7.2 Ad Hoc Event Detection Submission Systems

We submitted an identical set of systems for the ad hoc task and pre-specified event detection. The table below shows the MAP scores for the different training conditions.

	FullSys	ASRSys	AudioSys	OCRSys	VisualSys
EK100	32.2%	8.0%	15.1%	5.3%	23.4%
EK10	14.3%	4.1%	5.8%	2.3%	10.8%
EK0	8.1%	2.5%	0.6%	3.0%	5.0%

Table 5: MED 2013 Ad Hoc Results

The performance of the different submissions were consistent with the trends observed in pre-specified, and was also competitive with the other submissions. Our EK10 and EK0 full systems were 2nd, while the EK100 was 3rd in terms of MAP. Further, our EK0 system had consistent performance between pre-specified and ad hoc demonstrating the generality of our concept detectors.

7.3 MER Submission

For the TRECVID Multimedia Event Recounting (MER) Task, we submitted a three-phase system that (1) detected concept instances from various modalities; (2) aggregated these detections by modality, filtering out detections with low confidence or low relevance to the event type at hand; and (3) generated a human-readable recounting containing itemized detections along with confidence and relevance information. The system combined concept detections from the following systems:

- **Audio-Visual Concepts:** We obtained these concepts using the system described in Section 3. For each test video, we applied all our concept detectors and pruned those concepts that had confidence below the threshold learned during training.
- **Automatic Speech Recognition (ASR):** We applied BBN’s ASR system on the audio stream, and then detected salient keywords in the speech transcript. We then included these keywords, as well as the start and end times of their utterances in our MER submission.
- **Videotext:** We applied BBN’s Videotext detection and recognition system on the videos and included the output in our MER submission.

Our submission had an accuracy of 64.96%, percent recounting review time (PRRT) of 50.59% and observation text precision of 1.78. Our system accuracy was 2nd across all submissions and was best among systems with PRRT<100%; i.e., systems for which viewing the evidence took less time than viewing the full video.

8 Acknowledgments

Supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center contract number D11PC20071. The U.S. Government is authorized to reproduce and distribute reprints for Government 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.

References

- [Smeaton et al. 2006] A. F. Smeaton, P. Over, and W. Kraaij, “Evaluation campaigns and TRECVID,” in *Proc. of the 8th ACM International Workshop on Multimedia Information Retrieval*, 2006.
- [Over et al. 2013] P. Over, G. Awad, M. Michel, J. Fiscus, G. Sanders, W. Kraaij, A. F. Smeaton, and G. Quénot, “TRECVID 2013 - An Overview of the Goals, Tasks, Data, Evaluation Mechanisms and Metrics,” in *Proc. of TRECVID 2013*, 2013.
- [Csurka et al. 2004] G. Csurka, C. Dance, L.X. Fan, J. Willamowski, and C. Bray, “Visual Categorization with Bags of Keypoints,” in *Proc. of ECCV International Workshop on Statistical Learning in Computer Vision*, 2004.
- [Lazebnik et al. 2006] S. Lazebnik, C. Schmid, and J. Ponce, “Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories,” in *Proc. of CVPR*, 2006.
- [Laptev et al. 2008] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, “Learning Realistic Human Actions from Movies,” in *Proc. of CVPR*, 2008.
- [Jiang et al. 2010] Y.-G. Jiang, X. Zeng, G. Ye, S. Bhattacharya, D. Ellis, M. Shah, and S.-F. Chang, “Columbia-UCF TRECVID 2010 Multimedia Event Detection: Combining Multiple Modalities, Contextual Concepts, and Temporal Matching,” in *Proc. of TRECVID 2010*, 2010.
- [Lowe 2004] D. Lowe, “Distinctive image features from scale-invariant keypoints,” *International Journal on Computer Vision*, vol. 60, no. 2, pp. 91-110, 2004.

- [Mikolajczyk et al. 2004] K. Mikolajczyk and C. Schmid, "Scale and affine invariant interest point detectors," *International Journal of Computer Vision*, vol. 60, pp. 63–86, 2004.
- [Laptev 2005] I. Laptev, "On space-time interest points," *International Journal of Computer Vision*, vol. 64, pp. 107–123, 2005.
- [Liu 2011] J. Liu, M. Shah, B. Kuipers, and S. Savarese, "Cross-View Action Recognition via View Knowledge Transfer," in *Proc. of CVPR*, 2011.
- [Pan et al. 2010] S. J. Pan, X. Ni, J.-T. Sun, Q. Yang, and Z. Chen, "Cross-domain sentiment classification via spectral feature alignment," in *Proc. of the International Conference on World Wide Web*, 2010.
- [Torresani et al. 2010] L. Torresani, M. Szummer, A. Fitzgibbon, "Efficient Object Category Recognition Using Classemes," in *Proc. of ECCV*, 2010.
- [van de Sande et al. 2010] K. E. A. van de Sande, T. Gevers and C. G. M. Snoek, "Evaluating Color Descriptors for Object and Scene Recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1582-1596, 2010.
- [Boureau et al. 2010] Y. Boureau, F. Bach, Y. Le Cun, and J. Ponce, "Learning mid-level features for recognition," in *Proc. of CVPR*, pp. 2559-2566, 2010.
- [Bay et al. 2008] H. Bay, A. Ess, T. Tuytelaars, and L. V. Gool, "Surf: Speeded up robust features," in *CVIU*, vol. 110, no. 3, pp. 346-359, 2008.
- [Chandrasekhar et al. 2011] V. Chandrasekhar, G. Takacs, D. Chen, S. Tsai, Y. Reznik, R. Grzeszczuk, and B. Girod, "Compressed histogram of gradients: a low bitrate descriptor," *International Journal on Computer Vision*, vol. 94, no. 5, 2011.
- [Felzenszwalb et al. 2010] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan, "Object Detection with Discriminatively Trained Part Based Models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, 2010.
- [Viswanathan et al. 2010] S. V. N. Vishwanathan, Zhaonan Sun, Nawanol Theera-Ampornpant, and Manik Varma, "Multiple Kernel Learning and the SMO Algorithm," in *NIPS*, vol. 22, pp. 2361-2369, 2010.
- [Natarajan et al. 2011] P. Natarajan, S. Tsakalidis, V. Manohar, R. Prasad, and P. Natarajan, "Unsupervised Audio Analysis for Categorizing Heterogeneous Consumer Domain Videos," in *Proc. of Interspeech*, 2011.
- [Natarajan et al. 2012] Pradeep Natarajan, Shuang Wu, Shiv Naga Prasad Vitaladevuni, Xiaodan Zhuang, Unsang Park, Rohit Prasad, Premkumar Natarajan, "Multi-channel Shape-Flow Kernel Descriptors for Robust Video Event Detection and Retrieval," in *Proc. of ECCV*, vol. 2, pp. 301-314, 2012.
- [Vitaladevuni et al. 2011] S. Vitaladevuni, P. Natarajan, R. Prasad, and P. Natarajan, "Efficient Orthogonal Matching Pursuit using sparse random projections for scene and video classification," in *Proc. of ICCV*, pp. 2312-2319, 2011.
- [Manohar et al. 2011] V. Manohar, S. Tsakalidis, P. Natarajan, R. Prasad, and P. Natarajan, "Audio-Visual Fusion Using Bayesian Model Combination for Web Video Retrieval," in *Proc. of the 19th ACM International Conference on Multimedia*, pp. 1537-1540, 2011.
- [Natarajan et al. 2011] P. Natarajan et al, "BBN VISER TRECVID 2011 Multimedia Event Detection System," in *Proc. of TRECVID 2011*, 2011.
- [Bo et al. 2010] L. Bo, X. Ren, and D. Fox, "Kernel descriptors for visual recognition," *NIPS*, pp. 244-252, 2010.
- [Bo et al. 2011] L. Bo, K. Lai, X. Ren, and D. Fox, "Object recognition with hierarchical kernel descriptors," in *Proc. of CVPR*, pp.1729-1736, 2011.
- [Heikkila et al. 2006] M. Heikkila, M. Pietikainen, and C. Schmid, "Description of interest regions with center-symmetric local binary patterns," *ICVGIP*, pp. 58-69, 2006.
- [Wang et al. 2013] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu, "Dense Trajectories and Motion Boundary Descriptors for Action Recognition," *International Journal of Computer Vision*, vol. 103, no. 1, pp. 60-79, 2013.
- [Li et al. 2010] L.-J. Li, H. Su, E. P. Xing, F.-F. Li, "Object Bank: A High-Level Image Representation for Scene Classification & Semantic Feature Sparsification," *NIPS*, pp. 1378-1386, 2010.
- [Patterson et al. 2012] G. Patterson and J. Hays, "SUN attribute database: Discovering, annotating, and recognizing scene attributes," in *Proc. of CVPR*, pp. 2751-2758, 2012.
- [Ng et al. 2012] T. Ng, B. Zhang, L. Nguyen, S. Matsoukas, X. Zhou, N. Mesgarani, K. Vesel, and P. Matjka, "Developing a speech activity detection system for the DARPA RATS program," in *Proc. of Interspeech*, 2012.
- [Over et al. 2011] P. Over, G. Awad, M. Michel, J. Fiscus, W. Kraaij, A. F. Smeaton, and G. Quénot, "TRECVID 2011 – An Overview of the Goals, Tasks, Data, Evaluation Mechanisms and Metrics," in *Proc. of TRECVID 2011*, 2011.
- [Peng et al. 2011] X. Peng, H. Cao, R. Prasad, and P. Natarajan, "Text extraction from video using conditional random fields," in *International Conference on Document Analysis and Recognition (ICDAR)*, pp. 1029-1033, 2011.

[Peng et al. 2013] X. Peng, H. Cao, S. Setlur, V. Govindaraju, and P. Natarajan, "Multilingual OCR research and applications: an overview," in *Proc. of the 4th International Workshop on Multilingual OCR*, vol. 1, pp. 1-8, 2013.

[Aly et al. 2012] R. Aly, K. McGuinness, S. Chen, N. E. O'Connor, K. Chatfield, O. Parkhi, R. Arandjelovic, A. Zisserman, B. Fernando, T. Tuytelaars, D. Oneata, M. Douze, J. Revaud, J. Schwenninger, D. Potapov, H. Wang, Z. Harchaoui, J. Verbeek, and C. Schmid, "AXES at TRECVID 2012: KIS, INS, and MED", in *Proc. of TRECVID 2012*, 2012.

[Bao et al. 2011] L. Bao, S. Yu, Z. Lan, A. Overwijk, Q. Jin, B. Langner, M. Garbus, S. Burger, F. Metze, and A. Hauptmann, "Informedia@TRECVID 2011," in *Proc. of TRECVID 2011*, 2011.