BBN VISER TRECVID 2014 Multimedia Event Detection and Multimedia Event Recounting Systems

Florian Luisier, Manasvi Tickoo, Walter Andrews
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

Ruslan Salakhutdinov
Department of Computer Science
University of Toronto, Toronto, Canada

Vlad Morariu, Larry Davis
Department of Computer Science
University of Maryland, US

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

In this paper, we describe the Raytheon BBN Technologies (BBN) led VISER system for the TRECVID 2014 Multimedia Event Detection (MED) and Recounting (MER) tasks. We present a comprehensive analysis of the different modules: (1) Metadata Generator (MG) – a large suite of audio-visual low-level and sematic features; a set of deep convolutional neural network (DCNN) features trained on the ImageNet dataset; automatic speech recognition (ASR); videotext detection and recognition (OCR). For the low-level features, we used D-SIFT, Opponent SIFT, dense trajectories (HOG+HOF+MBH), MFCC and Fisher Vector (FV) representation. For the semantic concepts, we have trained 1,800 weakly supervised concepts from the Research Set videos and a set of YouTube videos. These concepts include objects, actions, scenes, as well as noun-verb bigrams. We also consider the output layer of the DCNN as a 1,000-dimensional semantic feature. For the speech and videotext content, we leveraged rich confidence-weighted keywords and phrases obtained from the BBN ASR and OCR systems. (2) Event Query Generation (EQG) - linear SVM event models are trained for each feature and combined using probabilistic late fusion framework. Our system involves both SVM-based and query-based detections, to achieve superior performance despite the varying number of positive videos in the different training conditions. We present a thorough study and evaluation of different features used in our system. (3) Event Search (ES) - At search time, simple dot products with the SVM hyperplane are computed for each feature and consequently rescaled into a posterior probability score for each video. (4) Semantic Query Generation (SQG) - we use the Indri Document Retrieval System to search for the closest words/terms to the given event name in a static offline corpus of around 100,000 Wikipedia and Gigapedia documents. After basic text processing, these words (ranked by TF-IDF measure) are then projected into our concept vocabulary using a text corpus knowledge source (e.g., Gigawords). Further, to meet the tight timing constraints of this year's evaluations for all the above mentioned modules, we compressed all our features to 1 byte per dimension which significantly sped up the feature extraction, training and event search timings. Owing to optimized feature design, model training and strong compression scheme, we could drastically reduce the time spent for disk I/O during MG, EQG and ES.

Consistent with previous MED evaluations, low-level features still exhibit strong performance. However, their MED performance can now be matched by purely semantic features, even in the 100Ex and 010Ex training conditions. As a result, our modules stand amongst the fastest, while maintaining very strong performance. Our mean average precision (MAP) and Minimum Acceptable Recall (MR_0) results are consistently among the top 3 performers for all training conditions for both pre-specified and ad-hoc events. For the MER task, while 65% of the evaluators found the key evidence convincing, 75 % of the judges agreed with the query conciseness and logical appeal.

1 Introduction

Large volumes of multimedia videos are captured with surveillance cameras and cheap handheld digital devices. Such imagery can provide crucial information for identifying persons, objects and events of interest. The development of an adequately broad, fast and robust system for representing and recognizing multimedia events in large volumes of unconstrained imagery, by itself, would represent a major scientific breakthrough. To this effect, the annual TRECVID Multimedia Event Detection (MED) and Recounting (MER) evaluations [Over et al. 2014, Smeaton et al. 2006] aim to measure progress in developing such techniques, and strong performance has been reported in recent evaluations [Natarajan et al. 2013, Alt et al. 2013, Lan et al. 2013]. Although low-level features have been shown to perform well, their performance drops considerably when the training data is limited (e.g. 010Ex). Further, these features cannot be used directly in 000Ex and SQ conditions, where there is no positive visual exemplars to perform the traditional supervised learning paradigm.

In this paper, we provide an overview of the BBN's VISER system for the TRECVID 2014 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 1,800 weakly supervised concept detectors. For the concept model training, we use the weakly supervised learning framework [Wu et al. 2014] to train linear SVM [Fan et al. 2008] models for each of these concepts. In addition to last year, we also use a set of deep convolutional neural network (DCNN) [Jia et al. 2013] features trained on the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) dataset as a set of 1,000 dimensional semantic feature. For the

MED task, we fused these features with speech and videotext output using probabilistic late fusion to get the final system outputs. For the MER task, the outputs of the semantic concept detectors along with speech and videotext were combined to produce event recounting based on the SQ output. Lastly, in order to meet the timing constraints for MG, EQG and ES, we developed feature compression techniques for all the features to reduce the memory footprint and avoid heavy memory disk I/O incurred during the execution of the various modules. Our main findings can be summarized as follows:

- Low-level features continue to exhibit strong performance and form the core of our 100Ex and010Ex submissions;
- Although the low-level features remain strong, their MED performance can now be matched by purely semantic features, in both the 100Ex and 010Ex training conditions;
- Speech and videotext provide complementary evidences and consistently improve performance for both pre-specified and ad hoc event detection;
- Deep learning system achieves promising results, despite being just based on a single (visual) modality;
- Floating-point precision is unnecessary. The full precision Fisher vector can be discretized between 0 and 255. The two parameters of the affine rescaling must be stored in floating-point precision. At training and search time, the unsigned char feature is converted back to floating-point precision and rescaled to its original range of values. The reading/writing time is thus reduced by a factor of 4 without significant loss in performance;
- Our MER system provides robust evidence tagging and localization while reducing the key evidence viewing time to 43 % of the original video duration.

The rest of the paper is organized as follows. In Section 2, we describe our metadata generation (MG) system in detail. Section 3 describes our event query generation (EQG) system for 010Ex and 100Ex training conditions. In Sections 4 and 5, we describe our Semantic Query Generation (SQG) and Event Search (ES) systems respectively. We conclude with a discussion of experimental results in Section 6.

2 Metadata Generation (MG)

2.1 Low-level Features

We extracted and combined multiple low-level audio and visual features using Fisher Vector encoding [Peronnin et al. 2010]. Our low-level system combines a set of features that capture salient gradient, color, texture, motion and audio patterns. We considered the following features and systems:

Audio Features: We considered Mel-Frequency Cepstral Coefficient (MFCC).

Visual Features: We considered multiple visual features including a 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, a dense trajectory-based feature that combines shape and motion features (I-DT) [Wang et al. 2013].

We carefully engineered and studied the effects of model complexity vs. discriminativity trade-off. To achieve these goals, we undertook the following steps:

- Feature Normalization: Root Normalization [R. Arandjelovic et al. 2012] achieved the best classification results on our internal TRECVID validation dataset.

- Feature Sampling: The dense features were extracted spatially every 4 pixels at 5 different resolutions to increase scale invariance.

- Fisher Vector Encoding: We observed that reducing the number of GMM clusters while increasing the PCA dimension gave slightly better results for most low-level features.

2.2 Deep Learning

In addition, this year, we use a set of deep convolutional neural network (DCNN) features trained on the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) dataset. We have trained on a GPU machine an 8-layer deep convolutional neural network on the 1.2 million annotated images from the ILSVRC dataset. The output layer of the DCNN is used as a 1,000-dimensional semantic feature. We also include the last convolutional layer of the DCNN as a 4,096-dimensional mid-level feature for 010Ex and 100Ex training conditions. As can be seen below, the performance of the visual deep learning feature is very close to low-level features and audio-visual semantic features.

System	AP	R0	AUC
Visual Deep Learning	0.2797	0.5373	0.9611
Low-level	0.3718	0.5812	0.9742
Semantic	0.3980	0.6330	0.9813

Table 1: Performance comparison on BBN's Internal Video Test Partition (E021-E040). Low-level consists of the early fusion of D-SIFT, O-SIFT, I-DT and MFCC based Fisher Vectors. Semantic comprises of early fusion of the corresponding WSC features.

2.3 Video-Level Weakly Supervised Semantic Features (WSC)

Ability to detect high-level semantic concepts in videos is crucial for event recounting and event detection with a small set of positive training examples. However, there are several challenges in developing robust systems for detecting such semantic concepts. To this effect, 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. We have trained 1,800 weakly supervised concepts from the Research Set videos and a set of YouTube videos. We use linear SVM models for each of the concept models. These concepts include objects, actions, scenes, as well as noun-verb bigrams. We extracted the video-level concepts for D-SIFT, O-SIFT, MFCC and I-DT based fisher vector representations.

2.4 Language Features (ASR and OCR)

Like previous years, we observe that speech and videotext provide complementary evidences and consistently improve performance for both pre-specified and ad hoc event detection. To meet this year's timing constraints, we redesigned both the automatic speech recognition (ASR) and video text optical character recognition (OCR) pipelines. In contrast to our previous language content based metadata generation (CDR) pipelines, we have developed integrated pipeline from video input to metadata output for both speech and videotext. The integrated pipelines is ready for diverse parallel processing environments, with support for both single machine and computer cluster environments and significantly reduced network I/O. We have developed these pipelines for easier potential prototype integration.

2.4.1 Automatic Speech Recognition

We first use a 2-Gaussian GMM-based speech activity detection (SAD) [Ng. et al. 2012] applied on MFCC coefficients to automatically identify speech segments. Given the speech segments, we apply 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 developed a baseline speech recognition system using ~1600 hours of English broadcast news data, using upgraded tools that enable incorporation of ongoing and future advances in BBN's Byblos speech recognition systems. To accommodate the characteristics of the MED video data, we use updated dictionary and language model that result in about 2-4% (absolute) reduction in Word Error Rate, depending on different decoding settings. 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 12k. 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 develop a highly integrated speech pipeline optimized for efficient metadata generation. We streamline the processing steps, so network I/O is minimized and limited to only the input and the output. Specifically, all the processing for a single video happens on a single computer node, and only the final results are synchronized back to the disk RAID.

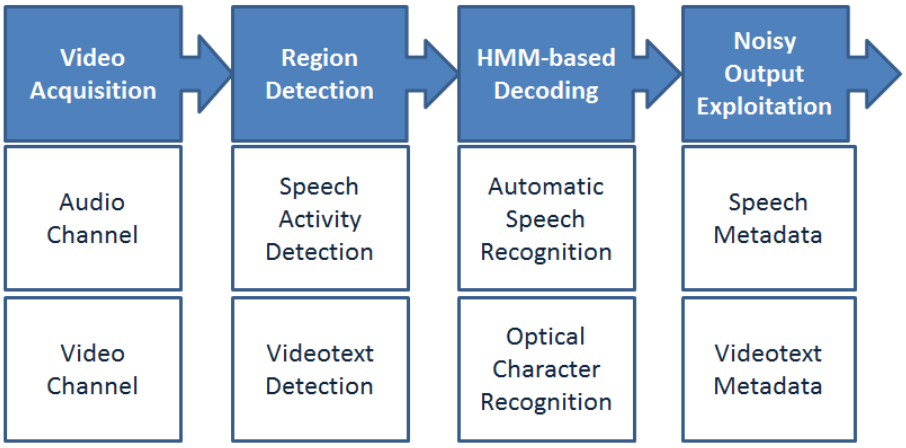

Figure 1: Flow Chart of the combined ASR and videotext OCR pipelines

2.4.2 Videotext OCR

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. First, MSER based videotext detector detects the bounding boxes. We leverage a statistically trained videotext detector based on SVM to estimate videotext bounding boxes. This content is then recognized by HMM-based multi-pass large vocabulary OCR. Similar to the ASR system, word lattices are used to encode alternative hypotheses.

The proposed system comprises of three steps (a) text localization (b) text line aggregation (c) text line recognition. We use MSER regions [Jain et al. 2014, Peng et al. 2011] as candidates and instead of using rules or geometric based grouping, we apply a text/non-text classifier over each candidate. We compute rich shape descriptors and compress them to very few dimensions while preserving discriminability using Partial Least Squares (PLS) technique. PLS technique enables use of a large set of features in classification, and speeds up the classification significantly. Each positively labeled candidate serves as an anchor region around which we group candidate regions based on geometric and color properties. At this step, we allow negatively labeled candidates to take part in text line aggregation, to overcome mistakes in the classification step. We binarize the detected text regions and pass it to the BBN HMM-based OCR system [Peng et al. 2011] for word recognition.

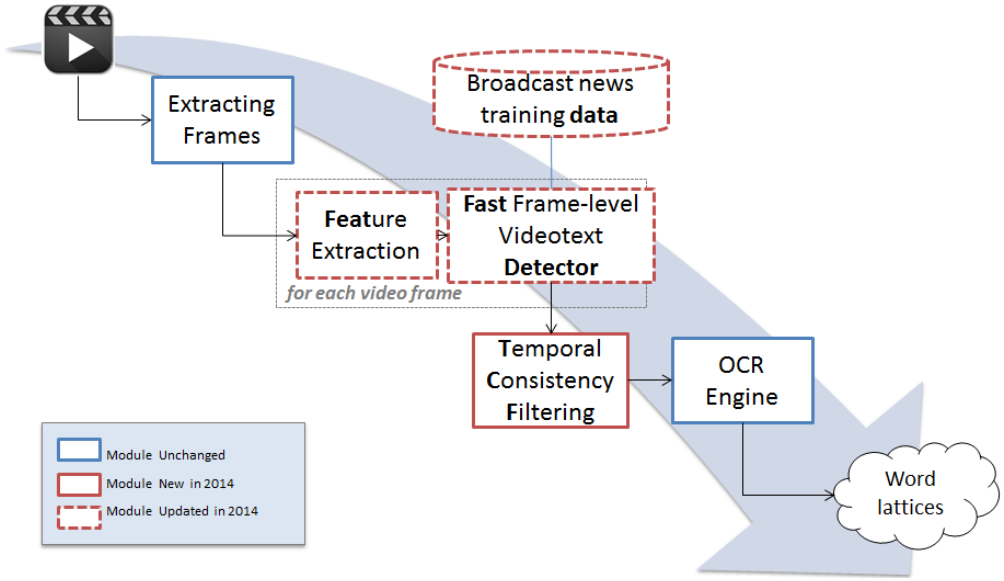

Figure 2: Flow Chart of the videotext detection and OCR pipeline

2.4.3 Language Representation for ASR and OCR

We start with a simple histogram representation built on the full lattice vocabulary after removing stopwords, which we denote W . This representation, $h_v = \{C(w)|w \in W\}$ is simply a histogram, i.e. the expected counts $C(w)$ for each word w for video v. The

vector h_v is l1 normalized to produce a normalized histogram $\widehat{h_v}$.
Since the lattice vocabulary is quite large and inclusive, we reduce the size of the representation by removing both extremely rare words in the vocabulary which are too infrequent to be robustly processed by machine learning techniques, as well as common words which are not useful in distinguishing videos of different events. Removing these extrema also help to offset noise from falsely detected speech and video text content. To this end, we compute the posterior-weighted document frequency df(w) of each word $w \in W$ over a sample collection of videos V as $df(w) = \sum_{v \in V} \min(1, C(w))$.

We then remove the 200 most common words as ranked by df, producing an upper cutoff $\overline{df}$. We also set a lower cutoff $\underline{df}$ by removing words with frequencies less than 0.001 times $\overline{df}$. These two cutoffs produce the compact vocabulary $W' = \{w | w \in W, \underline{df} < df(w) < \overline{df}\}$.

In our experiments, these cutoffs remove approximately 1/3 of the words in W for both ASR and OCR.
Using the shortened vocabulary W', we have $f_v = \{C(w) | w \in W'\}$.
Rather than l1 normalizing f_v, we normalize by the l1 of h_v, i.e., by the sum of posteriors of all words in the lattice of the given video, to produce $\tilde{f}_v$. This way, we encode in the normalization term information about the total amount of hypothesized language content in video v, to discount the potential false alarm in video text detection and speech detection.
We can further encode information about the potential classification power of each word in our vocabulary W' by weighting them by a revised inverse document frequency, $idf(w) = \log\left(\frac{\overline{df}}{df(w)}\right)$
We then define $g_v = \{C(w) \times idf(w) | w \in W'\}$.

We produce two representations by two different normalization schemes for gv: first, with a simple l1 normalization to produce an idf-weighted histogram $\widehat{g_v}$, and second, with the l1 norm of h_v, as we did with $\tilde{f}_v$, to produce $\widetilde{g_v}$.

Language Source	Representation	100Ex (MAP)	010Ex (MAP)
ASR	$\hat{h}_v$ (full histogram)	11.01%	3.93%
	$\tilde{f}_v$ (reduced vocab, full ℓ_1 normalization)	12.04%	6.68%
	$\hat{g}_v$ (reduced vocab + idf)	10.28%	3.70%
	$\tilde{g}_v$ (reduced vocab + idf, full ℓ_1 normalization)	12.38%	6.00%
OCR	$\hat{h}_v$ (full histogram)	6.47%	2.16%
	$\tilde{f}_v$ (reduced vocab, full ℓ_1 normalization)	8.26%	2.81%
	$\hat{g}_v$ (reduced vocab + idf)	6.75%	2.18%
	$\tilde{g}_v$ (reduced vocab + idf, full ℓ_1 normalization)	7.84%	3.25%

Table 2: Comparison of various language representations on BBN's internal test partition (E021-E040)

2.5 Analysis and Comparison

Thanks to the above mentioned efforts on features size reduction, our entire 2014 metadata store only takes around 100G of disk space for about 200,000 videos. As a comparison, our 2013 metadata store required about 1T of disk space for about 100,000 videos.

To summarize, the proposed optimization consists of two main changes:

1. Most computations are performed in floating point format. Only computations where numerical precision is crucial are performed in double format.
2. For a given video, all the features were extracted and encoded sequentially. We have now switched to a multi-threaded feature extraction and encoding strategy.

Table 3 reports the computation times (in seconds on a single Intel(R) Xeon(R) CPU E5450 @ 3.00GHz 8-core Linux machine) required to extract D-SIFT features, while Table 4 reports the computations times required for the Fisher vector encoding of the extracted D-SIFT features (both in seconds). As observed, our optimization efforts have drastically reduced last year's computation time, while maintaining a similar classification performance.

Table 3: D-SIFT feature extraction timing comparison between BBNVISER's 2013 and 2014 systems.

Number of Descriptors	2013 System [Natarjan et al. 2013]	2014 System
35,112	34.93 sec	1.78 sec
215, 061	209.04 sec	10.19 sec
669, 900	977.10 sec	41.16 sec

Number of Descriptors	2013 System [Natarajan et al. 2013]	2014 System
35,112	12.12 sec	1.31 sec
215, 061	70.46 sec	6.87 sec
669, 900	223.97 sec	20.64 sec

Table 4: D-SIFT Fisher Vector Encoding timing comparison between BBNVISER's 2013 and 2014 systems

We observed that the full precision Fisher vector can be rescaled between 0 and 255 and stored as unsigned char values without much loss of performance. The two parameters of the affine rescaling must be stored in floating-point precision. At training and search time, the unsigned char feature is converted back to floating-point precision and rescaled to its original range of values. The reading/writing time is thus reduced by a factor of 4 without loss in performance. The table below compares the memory footprint between the 2013 [Natarajan et al. 2013] and 2014 systems.

	2013 System		2014 System	
Feature Type	Count	Total Size per Video (KB)	Count	Total Size per Video (KB)
Appearance	2	2,097	1	97
Color	1	2,097	1	65
Motion	3	6,291	1	100
Audio	3	655	1	46
Deep Learning	0	N/A	2	26
Semantic	6	6	9	24
Language	2	176	2	28
SUM	17	11,322	17	386

Table 5: Comparison of features used in BBNVISER's 2013 and 2014 systems

3 Event Query Generation (EQG)

Besides a drastic compression of the metadata store, our 2014 system only involves linear classifiers, whereas our 2013 system relied on nonlinear kernel SVM classifiers. As a result, it takes only 10-15 min to train each of our 2014 event detectors on a COTS machine, while it took a few hours to train our 2013 event detectors on a cluster of computer nodes. Finally, it takes less than 5 min to search a database of 200,000 videos on a COTS machine with our 2014 system. The same search would have taken hours with our 2013 system due to the use of nonlinear SVMs. The table below compares the performance of SIFT Fisher vectors with those from last year. We can see the strong performance this year despite drastic dimensionality reduction and the use of linear SVMs.

Feature	**Dimensionality**	**Classifier**	**MAP**	**MR0**	**AUC**
D-SIFT 2013	262,144	RBF SVM	0.2589	0.3530	0.9271
Opp-D-SIFT 2013	524,288	RBF SVM	0.2620	0.3789	0.9327

D-SIFT 2014	131,072	Linear SVM	0.2857	0.3605	0.9307
Opp-D-SIFT 2014	262,144	Linear SVM	0.2809	0.3960	0.9317

Table 6: Comparison of D-SIFT and O-SIFT features on BBN's internal test partition for 100Ex training condition. Note that the features this year are stronger despite much lower dimensionality

While linear SVM perform well for low-level features, we observe that for semantic features, non-linear SVMs perform better than linear SVMs .Table 7 shows the comparison on BBN's internal test partition.

Kernel Type	**MAP**	**Test Time (sec/100 videos)**
Linear	0.2451	**0.08**
Intersect (Non-Linear)	**0.3071**	96.0

Table 7: Comparison of linear v/s non-linear kernels for visual semantic features (100Ex)

To overcome this, we used a non-linear kernel approximation [Vedaldi et al. 2012] which uses a pre-computed linear feature mapping based on the Fourier sampling theorem to approximate homogenous, additive kernels such as chi^2, Intersect and Hellinger's. After the mapping, a standard linear SVM can be used. Such a mapping gives performance close to the original non-linear kernels while being an order of magnitude faster at the search time.

Kernel Type	MAP	Test Time (sec/100 videos)
Linear	0.2451	**0.08**
Intersect (Non-Linear)	**0.3071**	96.0
Approx. Intersect (Linear)	0.3059	0.40

Table 8: Non-linear kernel approximation improves performance without much increase in event search

For the 010 Ex and 100 Ex training conditions, the soft-margin linear SVM event models involve early fusion of features within a given modality and late fusion of the ASR, OCR and audio-visual features. The parameters of the SVM classifier are optimized using k-fold cross validation based on 100 (resp. 10) positives and about 5,000 negatives (event background). The EQG runs on a single 16-core COTS computer. It takes from 10min to 30min to train an event model, depending on the number of cross-validation folds and the sampling of the parameters space.

For the 000Ex, the EQG expands the SQ XML to the corresponding concept lexicon space (Sec 4) by adding more relevant concepts to generate an event-specific model separately for each of the feature vocabulary.

4 Semantic Query Generation (SQG)

The semantic query (SQ) is the translation of the user-defined event query into a system query. It is the only semantic representation of an event. When positive training examples are available (e.g., in the 100Ex and 010Ex training conditions), the semantic query can be augmented/modified based on the information learned at training time to form the final event query. In the SQ training condition, the semantic query is the only input to the system, while in the 000Ex training condition, about 5,000 background (unlabeled) videos are also provided. In our 2014 system, we generated the SQ automatically, as described below.

We use the event name (e.g., "birthday party") as the query to the INDRI document retrieval system [Lemur Project] to retrieve a ranked list of the most relevant documents for the given event query. In practice, we keep the top-15 documents. The INDRI text corpus consists of about 1 million general Wikipedia articles which have been downloaded into the system before any knowledge of the user query. We apply standard text processing techniques (e.g., stop word removal, lemmatization …) to the top-15 retrieved documents. Each word in the documents is then weighted by its Term Frequency (TF) measured across all the top documents. The words having the highest TF are grouped together into a vector which represents the event query.

Once the event query representation is obtained, we use a text expansion based method to project the query into our various semantic feature spaces (ASR, OCR, audio-visual semantics). We use the large text corpus Gigaword for this projection. Our semantic features consist of speech (ASR), text (OCR), weakly supervised audio-visual concepts (VideoWords) and ImageNet deep convolutional neural network (CNN) features. Once the expanded event query feature vector and video feature vector are computed, we use a cosine similarity measure to rank the videos for a given event. Finally, we fuse retrievals from each of the features and modalities using a simple linear combination to exploit the complementary nature of the different modalities and concept vocabularies.

For the SQ module, we use only the top 24 features for the language (ASR & OCR) and 9 (audio-visual and CNN) at the time of search, while for the 000Ex system we retained 30 language concepts and 15 audio-visual and DCNN concepts. The retained concepts for each event are returned in the SQ XML and 000Ex XML, respectively. The performance of our SQ and 000Ex system is reported in Table 3.

	MED'13 (E021-E030)		**MED'14 (E021-E040)**	
	SQ	**000Ex**	**SQ**	**000Ex**

MAP	0.0431	0.0643	0.0830	0.0901
MR0	0.1628	0.2183	0.2322	0.2502
AUC	0.8670	0.9175	0.8869	0.9233

Table 9: Comparison of SQ, 000Ex training conditions on BBN's internal test partition (2013 and 2014 systems)

5 Event Search (ES)

Using the features described and discriminative event specific models, the ES module searches the video test database for a particular event and returns a list of videos ranked based on the corresponding probability scores of belonging to the searched event. It also returns a detection threshold above which a video is considered as a positive for the searched event. The score of each test video for a particular event is obtained by applying the event model to the test video. Since it only involves the computation of inner products and some posterior score normalization, the event search is very fast.

Late fusion of multiple sub-systems consistently improves performance of our 100Ex and 010Ex systems compared to the best single sub-system. Further, fusion of 000Ex systems with 010Ex produced additional gains.

For our 0-shot system, we combined WSC features from D-SIFT, O-SIFT, I-DT and MFCC features to obtain a score from the video-level concepts. This was then combined with scores from ASR, OCR and CNN modalities. As expected, the audio-visual semantic features had the strongest performance, followed by high precision OCR and ASR systems.

For the 010Ex and 100Ex test conditions, since the limiting condition for calculating the video level score is the dimensionality of the underlying feature space, we divide the test partition videos into blocks, each of which is then processed by a separate core on the COTS machine. Whereas, the 000Ex involves only the semantic features (roughly similar feature dimensionality), we perform the multi-threading at feature level. It takes about ~2min to rank a list of ~200,000 videos across all modalities on a single 16-core COTS computer. Based on these observations, we submitted our systems to NIST for the TRECVID MED pre-specified and ad hoc conditions.

6 Experiments and Results

In this section, we present results for the systems we submitted for the Pre-Specified and Ad Hoc MED tasks, and for MER, on the various training conditions.

6.1 Pre-Specified Event Detection Submission Systems

For the pre-specified task, we submitted systems for all the training conditions. The table below presents the MAP scores for the different submissions.

	MAP	**MR0**
100Ex	29.8%	56.3%
010Ex	18.0%	41.7%
000Ex	5.7%	24.3%
SQ	5.3%	20.3%

Table 10: MED 2014 Pre-Specified Results

As expected, there is a large drop in performance from EK100 to EK10 to EK0 and SQ only. Our pre-specified submissions overall had strong performance in comparison to other TRECVID 2014 submissions. Our system finished 2nd for 000Ex and SQ tasks and 3rd for 100Ex and 010Ex training conditions.

6.2 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.

	MAP	**MR0**
100Ex	22.6%	46.9%
010Ex	10.9%	33.3%
000Ex	3.7%	14.7%

SQ	3.1%	11.7%

Table 11: MED 2014 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 SQ and 000Ex systems were 2nd, while the 100Ex and 010Ex systems were 3rd in terms of MAP. Further, our SQ and 000Ex systems had consistent performance between pre-specified and ad hoc demonstrating the generality and robustness of our concept detectors.

6.3 Metadata Generation Times

Our Metadata Generation module was one of the fastest compared to other performers. This was done while maintaining strong MAP performance. While we took only 0.027 hours processing time per hour of video to generate all the features (Sec 2) for ~1,500 videos, we took 0.262 hours ~200, 000 Eval videos.

6.4 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, based on the initial event-specific semantic query (Sec 4); 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 2.3. 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.

For query conciseness, our submission received 17% strongly agree (highest) and 59 % agree votes nominated by 5 human judges. For key evidence convincing, we had the lowest strongly disagree (lowest) votes (7 %) and the highest strongly agree votes (27 %) nominated by 5 judges.

7 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. 2014] P. Over, G. Awad, M. Michel, J. Fiscus, G. Sanders, W. Kraaij, A. F. Smeaton, and G. Quénot, "TRECVID 2014 - An Overview of the Goals, Tasks, Data, Evaluation Mechanisms and Metrics," in *Proc. of TRECVID 2014*, 2014.

[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.

[Natarajan et al. 2013] P. Natarajan, S. Wu, F. Luisier, X. Zhuang, M. Tickoo, G. Ye, D. Liu, S. Chang, I. Saleemi, M. Shah, V. Morariu, L. Davis, A. Gupta, I. Haritaoglu, S. Guler, A. Morde, "BBN VISER TRECVID 2013 Multimedia Event Detection and Multimedia Event Recounting Systems," in *Proc. of TRECVid 2013*, 2013.

[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. Mikoljczyk and C. Schmid, "Scale and affine invariant interest point detectors," *International Journal of Computer Vision*, vol. 60, pp. 63–86, 2004.

[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.

[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.

[Vedaldi et al. 2012] A. Vedaldi, A. Zisserman, "Sparse Kernel Approximations for Efficient Classification and Detection," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012

[Jia et al. 2013] T. Jia, "{Caffe}: An Open Source Convolutional Architecture for Fast Feature Embedding," http://caffe.berkeleyvision.org

[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.

[Lemur Project] http://www.lemurproject.org/

[Viswanathan et al. 2010] S. V. N. Vishwanathan, Zhaonan Sun, Nawanol Theera-Ampornpunt, and Manik Varma, "Multiple Kernel Learning and the SMO Algorithm," in *NIPS*, vol. 22, pp. 2361-2369, 2010.

[Peronnin et al. 2010] F. Perronnin, J. Sánchez, and T. Mensink. "Improving the Fisher kernel for large-scale image classification." In Proc. ECCV, 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.

[Fan et al. 2008] R. Fan, K. Chang, C. Hsieh, X. Wang, C. Lin, "LIBLINEAR : A Library for Large Linear Classification," in *Journal of Machine Learning Research (JMLR)*, 2008

[R. Arandjelovic et al. 2012] R. Arandjelović, A. Zisserman, "Three things everyone should know to improve object retrieval," in

IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2012

[Jain et al. 2014] Arpit Jain, Xujun Peng, Xiaodan Zhuang, Pradeep Natarajan, Huaigu Cao, "Text detection and recognition in natural scenes and consumer videos," in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1245-1249, 2014.

[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.

[Wu et al. 2014] S. Wu, S. Bondugula, F. Luisier, X. Zhuang, P. Natarajan "Zero-shot Event Detection using Multi-modal Fusion of Weakly Supervised Concepts," IEEE *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2665-2672, 2014.

[Wang et al. 2013] H. Wang, and C. Schmid "Action Recognition with Improved Trajectories," IEEE *International Conference on Computer Vision (ICCV)*, 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. 2013] R. Aly, R. Arandjelovic, K. Chatfield, M. Douze, B. Fernando, K. McGuinness, N. E. O'Connor, D. Oneata, O. Parkhi, D. Potapov, J. Revaud, C. Schmid, J. Schwenninger, D. Scott, T.Tuytelaars, J. Verbeek, H. Wang, and A. Zisserman, "AXES at TRECVID 2013", in *Proc. of TRECVid 2013*, 2013.

[Lan et al. 2013] Z. Lan, L. Jiang, S. Yu, C. Gao, S. Rawat, Y. Cai, S. Xu, H. Shen, X. Li, Y. Wang, W. Sze, Y. Yan, Z. Ma, N. Ballas, D. Meng, W. Tong, Y. Yang, S. Burger, F. Metze, R. Singh, B. Raj, R. Stern, T. Mitamura, E. Nyberg, and A. Hauptmann "Informedia@TRECVID 2013," in *Proc. of TRECVid 2013*, 2013.