BBN VISER TRECVID 2012 Multimedia Event Detection and Multimedia Event Recounting Systems

Pradeep Natarajan, Prem Natarajan, Shuang Wu, Xiaodan Zhuang, Amelio Vazquez-Reina, Shiv N. Vitaladevuni, Kleovoulus Tsourides, Carl Andersen, Rohit Prasad
Speech, Language, and Multimedia Business Unit, Raytheon BBN Technologies, Cambridge, 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

Yue Ng, Brandyn White, Larry Davis
Department of Computer Science
University of Maryland, College Park, USA

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

Ismail Haritaoglu
Polar Rain Inc.
Menlo Park, USA

Abstract

We describe the Raytheon BBN Technologies (BBN) led VISER system for the TRECVID 2012 Multimedia Event Detection (MED) and Recounting (MER) tasks. We present a comprehensive analysis of the different modules in our evaluation system that includes: (1) a large suite of visual, audio and multimodal low-level features, (2) modules to detect semantic scene/action/object concepts over the entire video and within short temporal spans, (3) automatic speech recognition (ASR), and (4) videotext detection and recognition (OCR). For the low-level features we used multiple static, motion, color, and audio features previously considered in literature as well as a set of novel, fast kernel based feature descriptors developed recently by BBN. For the semantic concept detection systems, we leveraged BBN's natural language processing (NLP) technologies to automatically analyze and identify salient concepts from short textual descriptions of videos and frames. Then, we trained detectors for these concepts using visual and audio features. The semantic concept based systems enable rich description of video content for event recounting (MER). The video level concepts have the most coverage and can provide robust concept detections on most videos. Segment level concepts are less robust, but can provide sequence information that enriches recounting. Object detection, ASR and OCR are sporadic in occurrence but have high precision and improves quality of the recounting. For the MED task, we combined these different streams using multiple early/feature level and late/score level fusion strategies. We present a rigorous analysis of each of these subsystems and the impact of different fusion strategies. In particular, we present a thorough study of different semantic feature based systems compared to low-level feature based systems considered in most MED systems. Consistent with previous MED evaluations, low-level features exhibit strong performance. Further, semantic feature based systems have comparable performance to the low-level system, and produce gains in fusion. Overall, BBN's primary submission has an average missed detection rate of 29.6% with a false alarm rate of 2.6%. One of BBN's contrastive runs has <50% missed detection and <4% false alarm rates for all twenty events.

Description of Submitted Runs

Pre-Specified Task:

BBNVISER_Baseline_2: This is our primary EKFull system. It uses a combination of 18 complementary low-level features, including grayscale, color, motion, audio, gradient-based and optical flow. We also convert ASR and OCR recognition results to feature data. Finally, we use video- and segment- level concept detectors trained using NIST-provided event kit descriptions and BBN annotations; these concept detections become a further source of feature data. We combine these features using an early fusion strategy. The threshold is tuned to minimize missed detections while meeting the Year 2 false alarm ceiling (4%).

BBNVISER_AltThresh_2: Our secondary EKFull system. In addition to the sub-systems used in BBNVISER_Fusion1, we use multimodal words-based features and classeme-based concept detections. All features are combined using a late fusion strategy that uses the same tuned threshold used by BBNVISER_Fusion1.

BBNVISER_ EK10Ex-lateDBG_3: This is our submission for EK10Ex. It combines the 18 low-level features and a keyword-based OCR system. Feature combination and thresholds are as in BBNVISER_Fusion2, but trained using the EK10Ex condition.

Ad Hoc Task:

BBNVISER_Baseline_4: This is our primary EKFull system. It uses the 18 low-level features as well as ASR and OCR information. A late fusion strategy is used.

BBNVISER_EK10ExFusionNRNew_4: This is our primary EK10Ex system. It is similar to the EKFull system, aside from being trained using the EK10Ex condition. It uses only the true positives for each event specified in the EK10Ex partition for training.

BBN_EK10ExLLFeatFusionRNR_4: Our secondary EK10Ex system. It is similar to BBN_EK10Ex2 except that it uses both the true EK10Ex positives as well as the EK10Ex related videos.

1 Introduction

The ability to analyze large volumes of unconstrained web videos and detect events of interest has several compelling applications. Significant progress has been made in recent years in developing such capabilities. This is reflected in the strong performance reported in recent TRECVID evaluations [Natarajan et al. 11][Jiang et al. 2010]. 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 led VISER team's system for the TRECVID 2012 Multimedia Event Detection (MED) and Recounting (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 combined a suite of standard features, as well as a set of novel, fast, kernel-based feature descriptors developed by BBN. Our semantic system included video level concepts as well as those detected from short temporal segments in the video. 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 findings can be summarized as follows:

- As in previous years, low-level features continue to exhibit strong performance and form the core of our MED system.
- Kernel based feature descriptors allow definition of a rich set of pixel level features and provide significant performance improvements when combined with standard features.
- Semantic features produce small gains when combined with low-level features, but significantly improve recounting ability of the system.
- Speech and videotext are effective in retrieving videos containing such content. However, their occurrence is sporadic, especially for the MED 12 events.
- Kernel SVMs have strong performance in the EKFull condition, but linear classifiers have comparable or slightly better performance in the EK10Ex condition.
- We did not observe any significant performance degradations or challenges in going from pre-specified to the Ad Hoc event detection condition.

The rest of the paper is organized as follows – in Section 2 we describe our low-level feature system in detail. In Section 3, we describe 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

Combination of multiple, diverse, low-level audio and visual features have consistently shown strong performance in a range of video retrieval tasks (e.g. [Jiang et al. 2010][Natarajan et al. 2011]). Building on these results, we combined a large suite of low-level features in our system. In particular, we combined 4 classes of low-level features:

Appearance Features: These model local shape patterns by aggregating quantized gradient vectors in grayscale images. Our system included SURF [Bay et al. 2008], SIFT [Lowe 2004], D-SIFT [Boureau et al. 2010], CHoG [Chandrashekar et al. 2011].

Color Features: These features model color patterns. Our system included RGB-SIFT, O-SIFT and C-SIFT [van de Sande et al. 2010].

Motion Features: These are extracted from overlapping spatio-temporal patches in video and capture optical flow patterns. Our system included Spatio-temporal interest point based features (STIP) [Laptev 2005] and their dense version.

2.1 Kernel Descriptors

Despite their success, these standard features are hand-designed and aggregate image or flow gradients using a pre-specified, uniform set of orientation bins. Kernel descriptors [Bo et al. 2010] generalize such orientation histograms by defining match kernels over image patches, and have shown superior performance for visual object and scene recognition. In our work, we make two contributions: first, we extended kernel descriptors to the spatio-temporal domain to model salient flow, gradient and texture patterns in video. Further, based on prior studies in images [van Sande et al. 2010], we applied our kernel descriptors to extract features from different color channels. Second, we presented a fast algorithm for kernel descriptor computation of O(1) complexity for each pixel in each video patch, producing two orders of magnitude speedup over previous approaches of kernel descriptors [1][3] and other popular motion features such as STIP and HoG3D.

We extract kernel descriptors using image and optical flow gradients from a dense grid of spatio-temporal patches. For each of the features, we jointly model the image/motion features and spatial locations. In our system we included the following five features.

Grayscale Gradient Descriptors (Gray KDES-G): These features are extracted from the gradients L_x and L_y computed along the *x* and *y* directions, on gray scale image frames sampled from a video. We extract kernel descriptors for each patch.

Grayscale Gradient+Flow Descriptors (Gray KDES-FG): These features combine gradient and flow information, by concatenating the Gray KDES-G and Gray KDES-F descriptors at each video patch.

Grayscale Center Surround LBP Descriptors (Gray KDES-CL): The number of basis vectors for the joint LBP-position is 256X25=6400. Applying KPCA on the joint set results in too many components that slows down feature extraction, while using a fixed, small number of components (such as *200* in [Bo et al. 2010]) results in poor approximation of the space. In our work we used a variant of LBP called center-symmetric LBP (CS-LBP) [Heikkila et al 2006] that extracts a 4-dimensional binary vector from a pixel's neighborhood. This space has 16X25=400 dimensions and KPCA produces better approximation of the space with fewer principal components.

Color Gradient Descriptors (Color KDES-G): We first split the frame images sampled from video to constituent color planes. In our experiments we used the (*R*,*G*,*B*) planes, but we can use other color channels too. We then extract KDES-G features from the gradients computed on each plane. For each patch, we concatenate the KDES-G features from the different color planes.

Color Center Surround LBP Descriptors (Color KDES-CL): These features are similar to Gray KDES-CL, but are extracted from each of the (*R*,*G*,*B*) channels.

2.2 Joint Audio-Visual Bimodal Words

Joint audio-visual patterns often exist in videos and provide strong multi-modal cues for detecting events. In order to discover the visual-audio correlation, we apply a bipartite graph to model relation across the quantized words extracted from the visual and the audio modalities. We then apply graph partitioning to construct bi-modal words that reveal the joint patterns across modalities. In recent literature, bipartite graph has been used successfully in various applications [Liu et al. 2010][Pan et al. 2010]. To the best of our knowledge, this is the first work to apply bipartite graph to model the correlation between audio and visual codebooks. It offers several distinct advantages – the dimensionality of the features can be greatly reduced (from about 14,000 to 8,000) and the bi-modal words provide strong cues and discriminative power for detecting the MED’11 events.

First, we apply BoW to build audio words and visual words through the standard *k*-means clustering method separately. Then, a bipartite graph is constructed to capture certain relations (e.g. co-occurrence, causality, etc.) between the audio words and visual words. After that, the spectral clustering technique is used for graph partitioning. Finally, the original individual word in each modality (audio or visual) is re-quantized into the discovered bi-modal word clusters which are then used as the final feature.

To consider different scales of the temporal information, we apply two versions of the bi-modal technique: video-level and clip-level. For the video-level representation, the link relation in the bipartite graph is computed by measuring the co-occurrence statistics of audio and visual words in the entire video. For the clip-level, each video is first segmented into short clips and the link relation in the bipartite graph is computed from the co-occurrence statistics of audio and visual words in that clip only.

3 Semantic Audio-Visual 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 Concept Detection for Semantic Recounting

To address these 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 ~50,000 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 D-SIFT, KDES-FG, O-SIFT and MFCC features from all the videos. These features were chosen 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 area under curve (AUC) metric to ensure that the detected concepts have a reasonable level of accuracy. At the end of this process, we had a total of 750 video level concept detectors to capture salient visual and audio concepts.

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.

3.2 Segment-level Visual Concept Detection

In addition to the concepts described previously, we also developed detectors for concepts occurring in different video segments. For this we first collected annotations for frames captured from videos in the MED 12 training set. After we obtained these annotations, we analyzed the text to identify salient concepts using BBN's NLP tools, similar to the approach described in section 3.1. While scene detection in images is a widely studied problem in computer vision, frames extracted from videos have unique challenges compared to static images. Web videos have significant camera motion that makes the features extracted from a single frame noisy. Therefore, we aggregated features from short temporal segments in the video. We then aligned the concept annotations with these segments and then trained segment-level concept detectors based on D-SIFT, KDES-FG and O-SIFT features. After pruning concepts with few training examples or with concept detectors based on AUC, we obtained a total of 250 segment level concepts.

These concepts can be directly used for segment level recounting (MER). For MED, we applied the concept detectors in different segments of the video and the use the maximum of detection scores for each concept across all video segments. The vectors of these scores were then used to train event detectors.

3.3 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.

3.4 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 3000 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

Our approach for using the spoken language information in the audio track involved the following three modules. First, within the video clips, the speech segments were identified by a Speech Activity Detection (SAD) system. The SAD system employed two Gaussian mixture models (GMM), for speech and non-speech observations, respectively. Second, we applied BBN's large-vocabulary automatic speech recognition (ASR) system to the speech data. This system was adapted from a BBN ASR system trained on 1700-hour broadcast news. In particular, we adapted the lexicon and language model using MED descriptor files, relative web text data, and the small set of 101 video clips with annotated speech transcription. The acoustic models were adapted during ASR decoding for each video clip in an unsupervised fashion. Third, we used the distribution of predefined keywords within each video clip to leverage the hypothesized speech content for event detection. This bag-of-words feature representation was used in SVMs for each event.

Using the transcript, i.e., the 1-best output, from the ASR system may not be optimal for event detection. Noise, channel mismatch, domain mismatch may degrade the recognition performance and lead to high Word Error Rate (WER). Compared to the transcript, the ASR decoding lattice contains more alternatives and it is more likely to recover some of the keywords in the lattice. To evaluate whether a word in a lattice is an actual hit or a false alarm, we used the arc posterior probability of a word, as a soft count for that word. Given a video clip, we collected the soft counts from the lattices and create a histogram of keywords to represent the audio clip. To avoid the noise created by the words with very low posterior probability, we used a threshold to remove the words with very low posterior probabilities.

5 Videotext OCR

Given a video clip, videotext detection and recognition are applied on individual video frames. In particular, videotext detection first identified bounding boxes of videotext, and videotext recognition hypothesized videotext content and output videotext word lattices. We eliminated all the hypothesized special characters and retained only videotext that existed in bounding boxes with at least 75% area overlapping in two consecutive examined frames. Similar to the ASR approach, we leveraged the output decoding lattice for each hypothesized text bounding box, and used the lattice arc posteriors as the soft counts for different videotext words.

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 EKFull 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. For our EK10Ex systems, we used linear SVMs to train the individual subsystems since they outperformed kernel SVMs for this condition in our experiments.

6.2 System Combination

After training the different sub-systems and estimating their detection thresholds, we considered two possible strategies for combining the different sub-systems: Bayesian model combination (BAYCOM) and weighted average fusion, which are described in the following sub sections.

6.2.1 Bayesian Model Combination (BAYCOM)

In the first system combination approach, we combine different sub-system outputs using a Bayesian decision theoretic approach (BAYCOM). Let *M* be the number of models to combine, and r_i denote the output generated by model *i*. Here, r_i consists of the classification c_i generated by system *i*, along with the associated confidence score s_i, i.e. $r_i = (c_i, s_i)$. Let the event *c* mean “hypothesis *c* is correct” and *C* be the set of unique classes proposed by all systems. Then, the model selects the optimal hypothesis according to:

$$c^* = \underset{c \in C}{\operatorname{argmax}} P(c|r_1, \dots, r_M) \qquad (2)$$

In our system, we use class specific conditional probabilities and overcame data sparseness by smoothing the conditional probabilities with class independent probabilities. This year we modified the computation of the BAYCOM posterior as a likelihood ratio between the in-class and out-of-class probabilities, with each term incorporating prior distributions to prevent degenerative probabilities in the ratio. We also adapt our video specific weighting (see section 6.2.2 below) to weight the system conditional probabilities. These modifications produce smoother DET curves overall and alleviates the bimodal distribution output from the original BAYCOM formulation.

6.2.2 Weighted Average Fusion

In the second approach for combining different sub-systems, we used a variant of weighted average fusion where, 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_1), 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:

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

7 Experiments and Results

In this section, we will describe the different systems we submitted for the Pre-Specified and Ad Hoc MED tasks, and for MER.

Pre-Specified Event Detection Submission Systems: For our primary and secondary EKFull submissions, we fused our early fusion, 18 feature kernel SVM output together with the ASR, OCR, video level concept, and scene level concept classifiers using the weighted average fusion method we developed and successfully employed for last year’s MED 11 evaluations. Threshold estimation for our primary submission was performed through cross-validation on the training set, targeted to match the 12.5:1 missed detection to false alarm operating point specified for the task. Our secondary submission included the additional fusion of system outputs from Columbia, as well as a high-precision keyword-based retrieval system built on the OCR videotext. The small set of high confidence keywords for this videotext retrieval system were generated from the event kit descriptions, and while this system provides low recall, the precision of the retrieved videos is quite high. For our secondary submission, we also alternatively chose thresholds to retrieve a fixed number of videos on the progress set, aimed to ensure meeting the false alarm targets for this year (4% false alarm).

For our EK10Ex submission, we combined our early fusion system with the keyword-based OCR videotext retrieval system. We were unable to include SVM-trained ASR and OCR systems for EK10Ex, since the speech and videotext content were too sparse in the limited exemplar environment.

MED 12 Pre-Specified Results:

For BBN's primary MED 12 submission, missed detection and false alarm rates for all 20 events were within the second-year 50% P_{MD} and P_{FA} 4% box at the target error rate (TER); 18 of 20 events were within the box at the detection threshold. The system showed an average P_{MD} of 25.56% and P_{FA} of 2.64% across 20 events.

For BBN's contrastive submission, all 20 events were within the 50% P_{MD} 4% P_{FA} bounds at the detection threshold. The system achieved the same performance at TER. Its average P_{MD} was 20.97% and P_{FA} was 3.69% across 20 events.

BBN's EK10Ex system had 11 events within the second-year box at TER, and 6 events within the box at the detection threshold. It showed an average P_{MD} of 47.88% and P_{FA} of 4.44% across 20 events.

Ad Hoc Event Detection Submission Systems: We submitted 3 systems for the TRECVID Ad Hoc MED 12 evaluations.

Our primary Ad Hoc EKFull submission system comprises two main processing sub-systems: a subsystem that uses the Content Description Representation (CDR) Generator to create a CDR repository from the video corpus; and a subsystem that ingests user queries and uses the CDR Generator and the Event Agent Generator (EAG) to create event agents.

On the CDR front, we extracted several low-level appearance, color, motion and audio features from the video stream. We additionally combined these low-level features into several concept classifiers, as well as leveraging ASR and OCR information. We performed late fusion of all of these components to produce final system outputs.

Our EK10x fusion systems are similar to the EKFull system aside from being trained using the EK10Ex condition. The systems differ in their training protocols: the first system trained only using the true EK10Ex positives, while the second also trained upon the related videos in EK10Ex.

MED 12 Ad Hoc Results:

BBN's primary MED 12 EKFull Ad Hoc system meets the 75% P_{MD} 6% P_{FA} goals for all 5 events at both TER and the detection threshold. The system had an average P_{MD} of 22.22% and P_{FA} of 3.34% across 5 events.

For BBN's EK10Ex runs, BBNVISER_EK10ExFusionNRNew_4 met goals the 75% P_{MD} 6% P_{FA} for all 5 events, while BBNVISER_EK10ExLLFeatFusionRNR_4 met goals for 4/5 events at the detection threshold. The first system had P_{MD} of 46.86% and P_{FA} of 4.81%. The second system had P_{MD} of 46.00% and P_{FA} of 4.65%.

MER Submission: For the TRECVID Multimedia Event Recounting (MER) Task, we submitted a three-phase system that (1) first, 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) finally, 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 3.1. For each test video, we applied all our concept detectors and pruned those concepts that had confidence below the threshold learned during training.
- **Segment-level Concepts:** We obtained these concepts by splitting the video into segments and then applying the segment level concept detectors described in section 3.2. Again, we pruned those detections that had confidence below the learned threshold for that concept.
- **Person Detection:** Salient objects such as people are extremely useful in distinguishing between videos. To this end, we applied a part-based person detector from U. Chicago at different video frames. Further, we also learned color models to describe the upper and lower body clothing worn by the detected persons.
- **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.

MER 12 Results:

For Clip Identification Task, BBN's systems scored 62.22% correct on the Eval Set, and 58.89% correct on the Progress Set. For the Event Identification Task, the system scored 69.44% on the Eval Set and 82.2% on the Progress Set.

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

[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. CVPR*, 2006.

[Laptev et al. 2008] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning Realistic Human Actions from Movies," in *Proc. 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 TRECVID2010 Multimedia Event Detection: Combining Multiple Modalities, Contextual Concepts, and Temporal Matching," in *NIST TRECVID Workshop*, 2010.

[Lowe 2004] D. Lowe. Distinctive image features from scale-invariant keypoints. International Journal on Computer Vision, 60:91–110, 2004.

[Mikolajczyk et al. 2004] K. Mikoljczyk and C. Schmid. Scale and affine invariant interest point detectors. International Journal of Computer Vision, 60:63–86, 2004.

[Laptev 2005] I. Laptev. On space-time interest points. International Journal of Computer Vision, 64:107–123, 2005.

[Liu 2011] Jingen Liu, Mubarak Shah, Benjamin Kuipers, and Silvio Savarese, Cross-View Action Recognition via View Knowledge Transfer, IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), Colorado Springs, 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. international conference on World Wide Web, New York, NY, USA, 2010.

[Torresani et al. 2010] Lorenzo Torresani, Martin Szummer, Andrew Fitzgibbon. Efficient Object Category Recognition Using Classemes. European Conference on Computer Vision, 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, volume 32 (9), pages 1582-1596, 2010.

[Boureau et al. 2010] Y. Boureau, F. Bach, Y. Le Cun, and J. Ponce, Learning mid-level features for recognition. In CVPR, pages 2559-2566, 2010.

[Bay et al. 2008] H. Bay, A. Ess, T. Tuytelaars, and L. V. Gool. Surf: Speeded up robust features. CVIU, 110(3):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, May 2011.

[Felzenszwalb et al. 2010] P. Felzenszwalb, R. Girshick, D. McAllester, D. Ramanan. Object Detection with Discriminatively Trained Part Based Models. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 32, No. 9, September 2010

[Viswanathan et al. 2010] S. V. N. Vishwanathan, Zhaonan Sun, Nawanol Theera-Ampornpunt, and Manik Varma. Multiple Kernel Learning and the SMO Algorithm. In Advances in Neural Information Processing Systems 23, 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 Interspeech, Florence, Aug. 2011.

[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," To appear ICCV, Barcelona, 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," To appear ACM Multimedia, Scottsdale, AZ, Nov. 2011.

[Natarajan et al. 2011] P. Natarajan et al, "BBN VISER TRECVID 2011 Multimedia Event Detection System," TRECVID Workshop, Nov 2011.

[Bo et al. 2010] L. Bo, X. Ren, D. Fox, "Kernel descriptors for visual recognition," In: NIPS. (2010) 244-252

[Bo et al 2011] L. Bo, K. Lai, X. Ren, D. Fox, "Object recognition with hierarchical kernel descriptors," In: CVPR. (2011) 1729-1736

[Heikkila et al 2006] Heikkila, M., Pietikainen, M., Schmid, C.: *Description of interest regions with center-symmetric local binary patterns*. In: ICVGIP. (2006) 58-69