The 2013 SESAME Multimedia Event Detection and Recounting System

SRI International (SRI)	Robert C. Bolles, J. Brian Burns, James A. Herson, Gregory K. Myers, Stephanie Pancoast, Julien van Hout, Wen Wang, Julie Wong, Eric Yeh
University of Amsterdam (UvA)	Amirhossein Habibian, Dennis C. Koelma, Zhenyang Li, Masoud Mazloom, Silvia-Laura Pintea, Arnold W.M. Smeulders, Cees G.M. Snoek
University of Southern California (USC)	Sung Chun Lee, Ram Nevatia, Pramod Sharma, Chen Sun, Remi Trichet

ABSTRACT

The SESAME team submitted runs as a full participant in the MED13 evaluation, and submitted video, motion, and audio features; high-level semantic concepts for visual objects, scenes, persons, and actions; automatic speech recognition (ASR); and video optical character recognition (OCR). The individual types of features and concepts produced a total of eight event classifiers. We combined the event detection results for these classifiers using arithmetic mean and log-likelihood ratio fusion methods, and developed and applied a method for selecting the detection threshold. The SESAME system generated event recountings by selecting intervals based on the semantic concepts, and on concepts recognized by ASR and OCR. Our major findings are:

- Our strategy of first selecting the most informative interval for a video, and then determining the most appropriate event-related semantic concepts within that interval to display for multimedia event recounting (MER), produced the best ObsTextScore in the evaluation. (The ObsTextScore measures the judges' responses to the question *"How well does the text of this observation describe the snippet(s)?"*.)
- The multimedia event detection (MED) performance for 100Ex and 10Ex was dominated by the classifiers that exploited visual content.
- The ASR and OCR classifiers for 0Ex performed better than those trained with 10Ex.
- The log-likelihood ratio late-fusion method demonstrated improved performance over simple averaging of event detection scores for 100Ex, but not for 10Ex.

1. INTRODUCTION

The purpose of the MED13 evaluation [1] was to characterize the performance of multimedia event detection systems, which aim to detect user-defined events involving people in massive, continuously growing video collections, such as those found on the Internet. This is an extremely challenging problem, because the content of the videos in these collections is completely unconstrained, and the collections include varying qualities of user-generated videos, which are often made with handheld cameras and have jerky motions and wildly varying fields of view.

The goal of MER is to give users a human-understandable recounting for each clip that the MED system deems to be positive. Providing such evidence is not straightforward because humans usually think of an event in terms of specific associated semantic concepts, but the reliability of detectors for most individual semantic concepts is poor. The purpose of the MER evaluation was to assess the quality of recounting evidence associated with the MED retrieval results.

The SESAME team submitted runs as a full participant in the MED13 evaluation. These included runs under three training conditions: 100 positive examples (100Ex), 10 positive examples (10Ex), and no positive examples (0Ex). The SESAME team also submitted results for the MER evaluation.

Section 2 describes the SESAME MED system, Section 3 contains the results of the MED evaluation, and Section 4 describes the methods for MER and its evaluation.

2. SESAME MED SYSTEM DESCRIPTION

To handle this challenging problem, the SESAME MED system extracted a comprehensive set of heterogeneous low-level visual, audio, and motion features; high-level semantic concepts for visual objects, scenes, persons, and actions; and semantic concepts from the results of automatic speech recognition and video optical character recognition. Event detection scores for the individual types of features and concepts were generated by a total of eight event classifiers for the 100Ex and 10Ex training conditions. For 0Ex, the scores from four semantic-level classifiers (visual concepts, action concepts, ASR, and OCR) were combined. We combined the event detection results for these classifiers using arithmetic mean and log-likelihood ratio fusion methods, and developed and applied a method for selecting the detection threshold.

2.1 Visual Content

2.1.1 Visual Features and Concepts

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

Low-level features. We extracted low-level visual features for two frames per second from each video. We followed the bag-of-codes approach, which considers spatial sampling of points of interest, visual description of these points, and encoding of the descriptors into visual codes. For point sampling, we relied on dense sampling, with an interval distance of six pixels, and sampled at multiple scales. We used a spatial pyramid of 1x1 and 1x3 regions in our experiments. We used a mixture of SIFT, TSIFT, and C-SIFT descriptors [2]. We computed the descriptors around points obtained from dense sampling, and reduced them all to 80 dimensions with principal component analysis. We encoded the color descriptors with the aid of difference coding, using Fisher vectors with a Gaussian Mixture Model (GMM) codebook of 256 elements [3]. For efficient storage, we performed product quantization [4] on the features.

Semantic features. We detected semantic concepts for each frame using the low-level visual features per frame as input representation. We followed the approach in [5]. Our pool of detectors used the human-annotated training data from two publicly available resources: the TRECVID 2012 Semantic Indexing task [6] and the ImageNet Large-Scale Visual Recognition Challenge 2011 [7]. The former has annotations for 346 semantic concepts on 400,000 key frames from web videos. The latter has annotations for 1,000 semantic concepts on 1,300,000 photos. The categories are quite diverse and include concepts of various types; i.e., objects like *helicopter* and *harmonica*, scenes like *kitchen* and *hospital*, and actions like *greeting* and *swimming*. Leveraging the annotated data available in these datasets, together with a linear support vector machine (SVM), we trained 1,346 concept detectors in total. We then applied all available concept detectors to the extracted frames. After we concatenated the detector outputs, a concept vector represented each frame.

Three visual event classifiers. We included three visual event classifiers based on low-level and semantic features. To arrive at a video-level representation for the low-level visual event classifier, we relied on simple averaging. To handle imbalance in the number of positive versus negative training examples, we fixed the weights of the positive and negative classes by estimating the prior probabilities of the classes on training data. For classification, we used a linear kernel SVM. For the two video event classifiers based on semantic features**,** we aggregated the concept vectors per frame into a video-level representation. One approach used averaging and normalization, while the other approach used a new semantic encoding, which we will detail in the final notebook paper. On top of both concept representations per video, we used a non-linear SVM with χ^2 kernel with the same fixed weights to balance positive and negative classes.

2.1.2 Motion Features and Concepts

The low-level motion features were based on Dense Trajectories (DTs) [8] and MoSIFT [9]. We computed DT raw features with step size of 10 pixels and MoSIFT raw features with default parameters. The raw features were encoded using first- and second-order Fisher vector descriptors with a two-level spatial pyramid [10]. Descriptors were aggregated across each video. We generated four event classifiers: two with DT features using first- and second-order Fisher vector descriptors, and two with MoSIFT features using first- and second-order Fisher vector descriptors. The SVM with Gaussian kernel performed the classification for the MED13 task, whose parameters were selected using five-fold cross validation. Average fusion was chosen to combine the outputs from the same low-level feature.

Two event classifiers were generated based on action concept detectors. There are 96 action concepts annotated on the MED11 Event Kit provided by Sarnoff/UCF, and 101 action concepts from UCF 101 [11]. Sarnoff concepts contain actions that happen directly in MED videos, such as throwing, kissing, and animals eating. UCF 101 concepts are not directly relevant to MED videos – for example, cliff diving and playing violin – nonetheless, including these concepts still improves event detection scores. The action concept detectors were applied to small segments of videos and encoded by Hidden Markov Model Fisher vector descriptors [12]. The SVM with Gaussian kernel was used to train two event classifiers, one for each set of action concepts.

2.2 Audio Content

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

2.3 ASR

We ran an English ASR model trained on conversational telephone data and adapted to conversational data obtained in meetings. We performed supervised acoustic model adaptation using the LDC201208 release, and unsupervised adaptation using first-pass recognition. We also performed supervised and unsupervised language model adaptation to the ALADDIN domain. We used ASR to compute probabilistic word lattices, from which we extracted video-based one-gram word counts for MED, and local counts over 1-s intervals for MER. We then performed stemming to reduce the vocabulary size to about 40,000 words.

For the 100Ex and 10Ex conditions, the stemmed counts were mapped to features using a log-mapping; for each event, those features were used to train a linear SVM with an L1 penalty. For the 0Ex condition, the stemmed count for each word was then multiplied by a weight obtained by a non-linear mapping of the event profile for that word. By summing this quantity over all the words, we obtained the MED score for a given video. The non-linear mapping was a sigmoid that was tuned empirically. The event profiles were obtained from the Event Kit text by using term frequency–inverse document frequency (TF-IDF) weightings to rank the relevance of non-stopwords.

2.4 OCR

SRI's video OCR software detected and recognized text appearing in MED13 video imagery. This software recognizes both overlay text, such as captions that appear on broadcast news programs, and in-scene text on signs or vehicles [13]. The software was configured to recognize English language text.

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

For the 100Ex and 10Ex conditions, for each event, we trained a log-linear classifier regularized with dropout over the frequency histogram of words from the preprocessed video OCR text identified in the training videos. In addition to the preprocessing, we removed stopwords. At event detection time, the event detection score was computed from the likelihood of the video being a positive instance of the event, given the recognized keywords.

For the 0Ex condition, we used the same event profiles that were generated and used for ASR. The event detection score for each video was the cosine similarity between the word vector for the video and the word vector for the event profile.

2.5 Fusion

We applied two late fusion methods to combine the results from the various modalities: arithmetic mean and another method based on the log-likelihood ratio (LLR). Last year, we found that the arithmetic mean yielded results that were just as effective as results from more complex fusion models [14]. For arithmetic mean, the detection scores were normalized using a Gaussian function (i.e., computing the z score by removing the mean and scaling by the standard deviation). The Gaussian parameters were learned from the distribution of scores on a training set composed of Event Kit positives and the Event Background set.

We also experimented with the LLR fusion approach, which uses logistic regression to linearly combine the detection scores from various modalities into a value that best approximates the LLR of the hypothesis test for "Is this video a positive instance of the specified event?". Such a fusion approach was successfully applied in the past to other detection tasks, such as speaker identification [15] or spoken-term detection [16]. The LLR fusion approach computes different fusion weights for each modality and for each event. The training data for the fusion weights were the detection scores for each modality obtained by running ten-fold cross-validation with the Event Kit positives and the Event Background set. This procedure provided an unbiased detection score for each trial in the 10Ex and 100Ex conditions by training on 90% of the positives and negatives. For each trial, we created a feature vector by concatenating the scores of all of the N modalities. We included N indicator variables to account for the possibility of missing scores for some trials.

We then used logistic regression to train Fusion models. If the weight of a modality was found to be negative, the linear regression was recomputed with that modality removed. Since most modalities' detection scores were initially in [0,1], we mapped those to [-inf,+inf] using a logit mapping before fusion training. Since the mean average precision metric in the MED evaluation emphasizes the importance of correctly scoring the high-scoring detections, we found it beneficial to consider a modality's score to be missing if it fell below a threshold set to .002 in the posterior domain.

2.6 Thresholding

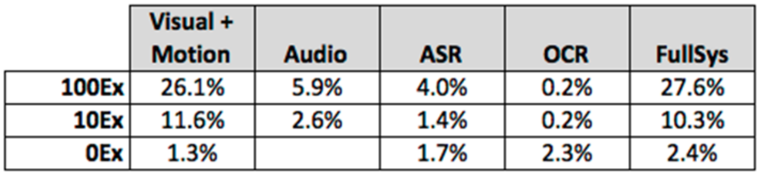
For the 100Ex and 10Ex conditions, we selected a detection threshold for each event using a set of approximately 4000 event trials with known truth data subsampled from the Research set and the Event Kits. We chose the threshold with the highest corresponding minimal acceptable recall metric, R_0, to be the optimal threshold.

When applying these thresholds to MEDTEST, we found that a single threshold for all events yielded slightly better R_0 values on average than event-specific thresholds. We believe that this was due to the small number of event positives (~20 – 40 per event for the 100Ex condition) in the training sets. In the future, we plan to use the full set of event positives in a cross-validation process to find event-specific thresholds that maximize R_0.

For the 0Ex condition, we estimated the thresholds based on the optimal thresholds for Events E001 through E005, which were part of the Research data set.

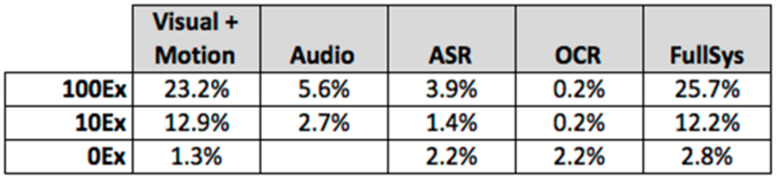
3. MED PERFORMANCE EVALUATION

Table 1: MED performance (mean average precision) on the PROGTEST data set for pre-specified events



	Visual + Motion	Audio	ASR	OCR	FullSys
100Ex	26.1%	5.9%	4.0%	0.2%	27.6%
10Ex	11.6%	2.6%	1.4%	0.2%	10.3%
0Ex	1.3%		1.7%	2.3%	2.4%

Table 2: MED performance (mean average precision) on the PROGTEST data set for ad hoc events



	Visual + Motion	Audio	ASR	OCR	FullSys
100Ex	23.2%	5.6%	3.9%	0.2%	25.7%
10Ex	12.9%	2.7%	1.4%	0.2%	12.2%
0Ex	1.3%		2.2%	2.2%	2.8%

Performance of the SESAME system was assessed on the PROGTEST and MEDTEST data sets. Tables 1 and 2 show MED performance (in terms of mean average precision) on the PROGTEST data set with the 100Ex, 10Ex, and 0Ex training conditions, and for five system configurations: ASR only, non-ASR audio, OCR only, non-OCR visual (which consisted of the visual features and concepts, and motion features and action concepts), and the full system (i.e., with all subsystems). Because the SESAME system does not currently include non-ASR audio concepts, there are no results for the AudioSys component for the 0Ex condition. Table 1 shows performance for pre-specified events, and Table 2 shows performance for ad hoc events.

Compared with the SESAME system description in Section 2, the SESAME system evaluated on the PROGTEST data set by NIST had two differences:

- A bug in the 100Ex and 10Ex event training for video OCR resulted in useless video OCR event detection scores.
- Instead of LLR fusion, the evaluated system used arithmetic mean fusion, because the results using the LLR method were not ready by the submission due date.

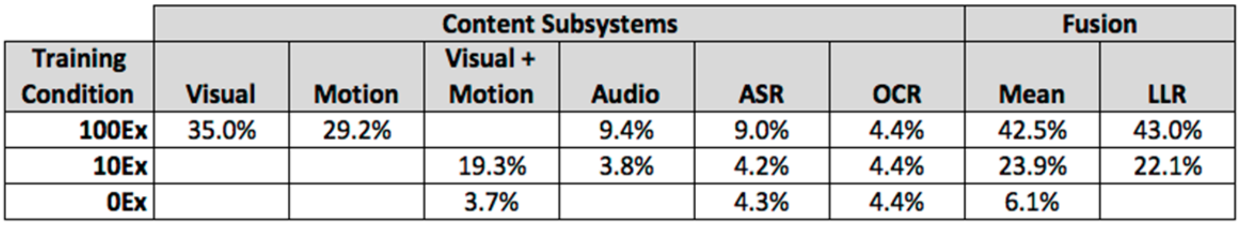
Tables 1 and 2 show that the MED performance for 100Ex and 10Ex was dominated by the classifiers that exploited visual content. In addition, the arithmetic mean fusion approach was not very effective: the performance of the full system was only slightly better than that of the visual system for 100Ex, and the performance of the full system was below that of the visual system for 10Ex. We believe that this was due to the noisy video OCR event detection scores.

The ASR classifier for 0Ex performed better than the one for 10Ex. We believe that this was due to the limited amount of training data available for 10Ex, and that building event detectors based on the Event Kit text resulted in better detection.

Table 3 shows the MED performance of the SESAME system after the event-training bug was fixed, and with LLR fusion on the MEDTEST data with the 100Ex, 10Ex, and 0Ex training conditions. The tables show the average precision of the event classifiers for the individual types of data and their fusion. The OCR event detection results for 0Ex performed better than those for 10Ex and 100Ex (not shown), so the 0Ex results were used in all of the fusion experiments on MEDTEST.

For 100Ex, the LLR fusion method performed slightly better than arithmetic mean. For 10Ex, the results from the visual and motion modalities were combined to limit the number of parameters in the LLR model. Still, LLR did not perform as well as arithmetic mean.

Table 3: MED performance on the MEDTEST data set



	Content Subsystems						Fusion	
Training Condition	Visual	Motion	Visual + Motion	Audio	ASR	OCR	Mean	LLR
100Ex	35.0%	29.2%		9.4%	9.0%	4.4%	42.5%	43.0%
10Ex			19.3%	3.8%	4.2%	4.4%	23.9%	22.1%
0Ex			3.7%		4.3%	4.4%	6.1%	

4. MULTIMEDIA EVENT RECOUNTING (MER)

Recounting consists of a concise textual summary of each piece of evidence and the source of the evidence: visible text (via OCR), video (not involving OCR), speech (transcribed via ASR), and/or non-speech audio (sounds not involving ASR textual transcription). For each piece of evidence, recounting also includes a confidence score; an importance score, indicating the important of the evidence in detecting the event; and spatial and temporal locations in the video where the evidence occurs.

4.1 MER Methods

The SESAME system generated event recountings for the 100Ex condition based on semantic concepts from the following multimedia sources: ASR, video OCR, and the 1551 automatically detected visual objects, scenes, persons, and actions. MER observations were selected from each source independently.

For the ASR and OCR MER data, the importance scores for individual text concepts were computed as a function of the weights of the text concepts in the MED classifiers. Up to four ASR results with importance scores above an event-specific, manually determined threshold were included in the MER results. For OCR, the text concepts with the top four importance scores were selected as candidates, and then filtered with an event-specific keyword list.

For the visual and action concepts, the video was divided into small segments, and concept detection scores were computed on each segment as part of the computation for MED. MER importance scores were computed for each segment in the video, based on an SVM classifier that was trained on the Event Kit and Event Background data sets. The segments with the four highest importance scores were selected for MER. The semantic concepts that contributed most to the importance score were selected as the MER observations for that segment.

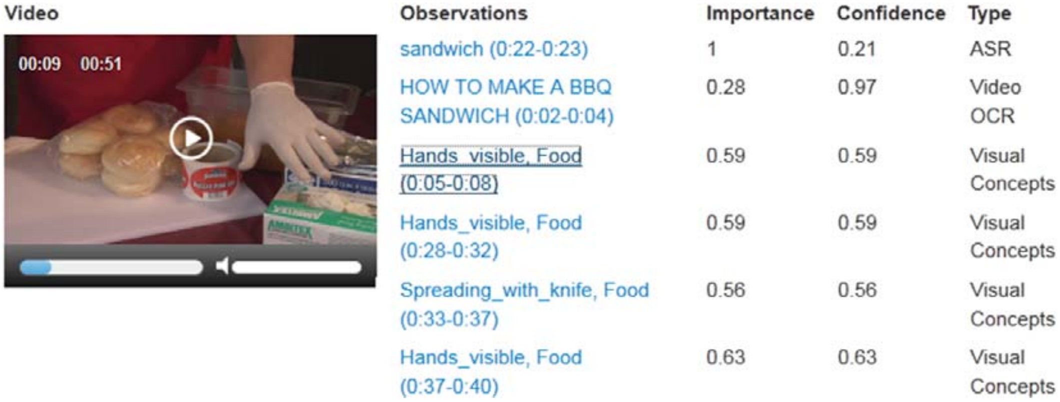
4.2 MER Evaluation

The purpose of MER is to help users accurately and rapidly assess whether each clip is truly a positive for the specified event. In the MER evaluation, the MER results were manually reviewed by NIST judges and were evaluated according to three measures:

- Accuracy: the degree to which the judges' assessments agreed with the MED ground truth
- Precision of the observation text: how well the text of the observations described the snippets
- Percent Recounting Review Time:: the percentage of clip time the judges took to perform the assessment

The SESAME MER results achieved an accuracy score of 64.1%, a precision score of 2.53 (out of a possible 4.0), and a Percent Recounting Review Time score of 41.83%.

Figure 1 shows a relatively successful example of MER produced by the SESAME system.



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

Figure 1. Successful example of MER for the event *Making a Sandwich*.

Although many of the visual concepts cited in the MER observations were not present in the video snippet, merely choosing the most relevant interval for viewing is often very helpful to the user. An informal manual examination of 250 video snippets generated from visual evidence indicated that approximately 75% contained enough content for a user to determine whether the video was an instance of the event. Therefore, the strategy of first selecting the best interval, and then determining the best semantic concepts within the interval to display for MER, seems to be worthwhile for visual and action concepts.

5. ACKNOWLEDGEMENT

This work was supported by the Intelligence Advanced Research Projects Activity (IARPA) via the Department of Interior National Business Center (DoI/NBC), contract number D11PC0067. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, or the U.S. Government.

6. REFERENCES

[1] Over, P. G. Awad, G., M. Michel, J. Fiscus, G. Sanders, B. Shaw, W. Kraaij, A.F. Smeaton, and G. Quéenot, TRECVID 2013 – An Overview of the Goals, Tasks, Data, Evaluation Mechanisms, and Metrics. *Proceedings of TRECVID 2013*. http://www-nlpir.nist.gov/projects/tvpubs/tv13.papers/tv13overview.pdf

[2] van de Sande, K.E.A., T. Gevers, and C. G. M. Snoek, Evaluating color descriptors for object and scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1582-1596, 2010.

[3] Sánchez, J., F. Perronnin, T. Mensink, and J. Verbeek, Image classification with the Fisher vector: Theory and practice. *International Journal of Computer Vision* 105:222-245, 2013.

[4] Jégou, H., M. Douze, and C. Schmid, Product quantization for nearest neighbor search. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(1):117-128, 2011.

[5] Habibian, A., K. E. A. van de Sande, and C. G. M. Snoek, Recommendations for Video Event Recognition Using Concept Vocabularies. *Proceedings of the ACM International Conference on Multimedia Retrieval*, Dallas, Texas, pp. 89-96, 2013.

[6] Over, P. G. Awad, G., M. Michel, J. Fiscus, G. Sanders, W. Kraaij, A.F. Smeaton, and G. Quéenot, TRECVID 2012 – An Introduction to the Goals, Tasks, Data, Evaluation Mechanisms, and Metrics. *Proceedings of TRECVID 2012*. http://www-nlpir.nist.gov/projects/tvpubs/tv12.papers/tv12overview.pdf

[7] Berg, A., J. Deng, S. Satheesh, H. Su, and F.-F. Li, ImageNet Large Scale Visual Recognition Challenge 2011. http://www.image-net.org/challenges/LSVRC/2011/

[8] Wang, H., A. Kläser, C. Schmid, and L. Cheng-Lin, Action Recognition by Dense Trajectories. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.

[9] Chen, M.-Y. and A. Hauptmann, *MoSIFT: Recognizing Human Actions in Surveillance Videos*. CMU-CS-09-161, Carnegie Mellon University, 2009.

[10] Sun, Chen and Ram Nevatia, Large-scale web video event classification by use of Fisher Vectors. *Workshop on the Applications of Computer Vision (WACV)*, pp. 15-22, 2013.

[11] Soomro, Khurram, Amir Roshan Zamir, and Mubarak Shah, UCF101: A Dataset of 101 Human Action Classes from Videos in the Wild. *Center for Research in Computer Vision*, CRCV-TR-12-01, November 2012.

[12] Sun, Chen and Ram Nevatia, ACTIVE: Activity Concept Transitions in Video Event Classification. *International Conference on Computer Vision* (ICCV), 2013.

[13] Myers, G., R. Bolles, Q.-T. Luong, J. Herson, and H. Aradhye, Rectification and recognition of text in 3-D scenes. *International Journal on Document Analysis and Recognition*, 7(2 3):147-158, July 2005.

[14] Myers, Gregory K., Ramesh Nallapati, Julien van Hout, Stephanie Pancoast, Ram Nevatia, Chen Sun, Amirhossein Habibian, Dennis C. Koelma, Koen E. A. van de Sande, Arnold W.M. Smeulders, and Cees G.M. Snoek, Evaluating Multimedia Features and Fusion for Example-based Event Detection. *Machine Vision and Applications*, July 2013.

[15] Brummer, N. and J. A. du Preez, Application-independent evaluation of speaker detection. *Computer Speech and Language*, 20(2-3):230-275, 2006.

[16] Akbacak, M., L. Burget, W. Wang, and J. van Hout, Rich system combination for keyword spotting in noisy and acoustically heterogeneous audio streams. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013.