The 2014 SESAME Multimedia Event Detection and Recounting System

SRI International (SRI)	Robert C. Bolles, J. Brian Burns, James A. Herson, Gregory K. Myers, Julien van Hout, Wen Wang, Julie Wong, Eric Yeh
University of Amsterdam (UvA)	Amirhossein Habibian, Dennis C. Koelma, Thomas Mensink, Arnold W.M. Smeulders, Cees G.M. Snoek
University of Southern California (USC)	Arnav Aggarwal, Song Cao, Kan Chen, Rama Kovvuri, Ram Nevatia, Pramod Sharma

1. ABSTRACT

The SESAME (video **SE**arch with **S**peed and **A**ccuracy for **M**ultimedia **E**vents) team submitted six runs as a full participant in the Multimedia Event Detection (MED) and Multimedia Event Recounting (MER) evaluations. The SESAME system combines low-level visual, audio, and motion features; high-level semantic concepts for visual objects, scenes, persons, sounds, and actions; automatic speech recognition (ASR); and video optical character recognition (OCR). These three types of features and five types of concepts were used in eight event classifiers. One of the event classifiers, VideoStory, is a new approach that exploits the relationship between semantic concepts and imagery in a large training corpus. The SESAME system uses a total of over 18,000 concepts. We combined the event-detection results for these classifiers using a log-likelihood ratio (LLR) late-fusion method, which uses logistic regression to learn combination weights for event-detection scores from multiple classifiers originating from different data types. The SESAME system generated event recountings based on visual and action concepts, and on concepts recognized by ASR and OCR. Training data included the MED Research dataset, ImageNet, a video dataset from YouTube, the UCF101 and HMDB51 action datasets, the NIST SIN dataset, and Wikipedia. The components that contributed most significantly to event-detection performance were the low- and high-level visual features, low-level motion features, and VideoStory. The LLR late-fusion method significantly improved performance over the best individual classifier for 100Ex and 010Ex. For the Semantic Query (SQ), equal fusion weights, instead of the LLR method, were used in fusion due to the absence of training data.

2. INTRODUCTION

The 2014 TRECVID MED and MER evaluations [1] characterize the performance of multimedia event detection systems, which find user-defined events involving people in massive and 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 user-generated videos in the collections are of varying quality. These videos are often made with handheld cameras and contain jerky motions and wildly varying fields of view.

The goal of MER is to give users a human-understandable recounting for each video that the MED system determines to be an instance of a user-defined event. 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 MED retrieval results.

The SESAME team submitted six runs as a full participant in this evaluation. These included both pre-specified (PS) and ad hoc (AH) runs under three training conditions: 100 positive examples (100Ex), 10 positive examples (010Ex), and SQ, which had no positive examples. A background dataset of 5000 negative examples was used for event training in the 100Ex and 010Ex runs. The SESAME team also submitted results for the MER evaluation.

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

3. SESAME MED SYSTEM DESCRIPTION AND EVALUATION

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, sounds, and actions; and semantic concepts from the results of ASR and video OCR. Event-detection scores for the individual types of features and concepts were generated by a total of eight event classifiers for the 100Ex and 010Ex training conditions. We combined the event-detection results for these classifiers using the LLR late-fusion method, and developed and applied a method for selecting the detection threshold. The SQ runs used the detection results of visual concepts, action concepts, ASR, and OCR.

The following sections describe the components of the SESAME system and the results of the MED evaluation tasks.

3.1 Visual Features and Concepts

To meet the computation-time requirements dictated by the evaluation schedule, we used two new approaches. The first approach, inspired by deep learning, used two types of proprietary features developed by Euvision: Visual High and Visual Low. The Visual High features are a set of 15,000 semantic concepts, and the Visual Low features are a vector of 4096 values. These features are based on 15,000 ImageNet category training examples. For both types of features, which were sampled one frame every two seconds, we trained an event classifier using a support vector machine (SVM) with a radial basis function (RBF) kernel. Compared to the scale-invariant feature transform (SIFT)-based, low-level visual features we used in 2013, the computation time for the new Euvision features decreased by almost an order of magnitude. Computing both the Visual Low and Visual High features for the MED14-Eval dataset took 200 hours on a single CPU, which is about 40 times faster than real time.

For the SQ runs, we selected about 10 concepts from the Event Kit text descriptions. However, not all of these concepts were within the capabilities of our detectors, so we had to find some correspondence between the two. We have a very large number of concepts, but they are typically very fine-grained and are not related to concepts in the queries. For example, the set of concepts may include a variety of birds, insects, reptiles, and dogs. While dogs are important in some queries, specific breeds of dogs may not be. To provide a better match between the 15,000 semantic concepts and query concepts, we developed a hierarchy of concepts. For example, “dog” became a parent node for a number of dog breeds. In the 010Ex test scenario, we examined which of the potential tags actually found support in training videos.

Our second approach built a different high-level representation on top of the Visual Low features. The VideoStory approach used these features at the video level and combined a VideoStory embedding [2] with an SVM classifier to find positive videos. This approach learned an embedding from videos and their semantic descriptions, which we obtained free of charge from the web with a simple spidering procedure.

We trained VideoStory on a collection of 46,000 YouTube videos and their titles. The embedding used a vocabulary of over 3000 semantic concepts.

More specifically, VideoStory learned a video embedding **W** to embed video features into VideoStories, and a story embedding **A** to back-project the VideoStories into their textual descriptions. In our 010Ex and 100Ex submissions, we used video embedding **W** to embed the event train and test videos into 1,000 dimensional VideoStory representations. We then generated the event query by training an SVM classifier with an RBF kernel with a fixed value of one (1) for C and gamma parameters. For SQ, the query was generated as a term vector that was extracted from the Event Kit descriptions. As an example, Figure 1 shows the VideoStory SQ terms for the event *Attempting a Bike Trick* (after stemming). During event search, the videos were represented as their predicted term vectors, which were obtained by applying the video embedding **W** followed by the story embedding **A**. The videos were then ranked based on the cosine similarity between their predicted term vectors and the event query.

activities	flip	object	slow
air	flipp	Obstacle	someth
attempt	floor	outdoor	sound
atv	forward	park	spinn
audience	ground	people	stand
audio	hand	person	steer
bicycle	handlebar	propel	stopp
bike	helmet	purpose	street
cheer	Hitt	ramp	surface
concrete	hold	rid	themselves
count	jump	ridden	to
definition	land	river	top
description	lot	scene	trick
difficult	made	seat	type
doe	motion	set	unicycle
dur	motorcycle	simply	vehicle
event	motoriz	sitt	wheel
feet	move	skate	

Figure 1. VideoStory SQ terms for event *"Attempting a Bike Trick,"* after stemming.

3.2 Motion Features and Concepts

Low-level motion features were generated using dense trajectory (DT) features coded by Fisher Vectors (FVs) [3]. Features were computed over two-second-long segments and over the entire video. For each segment, we computed scores for 164 action concepts by applying a linear SVM to the DTFV features. The action concepts were learned on the UCF101 action dataset [4, 5], the HMDB51 action dataset [6], and the NIST SIN dataset. An event classifier based on these features used a linear SVM learned from video-level DTFV features. A second event classifier, which used the relationships between action concepts in an Event Localization Model (ELM) [7], was developed but not used because there was not enough time in the MED/MER evaluation schedule to train the models and generate event-detection results.

3.3 Audio Features

For low-level audio features, we used FVs to extract a signature from modified mel-frequency cepstral coefficients (MFCCs). Classification was done using a linear SVM.

We extracted MFCCs from audio that was down-sampled to 8 kHz using a 25-ms analysis window and a 10-ms shift. We used a 256-dimensional Fast-Fourier Transform (FFT) that was passed through 40 linearly spaced sub-band filters, and obtained 20 cepstral coefficients after running a discrete cosine transform (DCT) on the log-energies. A 256-element Gaussian mixture model (GMM) trained with the MED Research dataset was used to derive mean and variance-based Fisher statistics. These statistics were L2-normalized and their square roots were used as feature values.

We developed an alternate approach to the basic deltas for contextualizing the MFCC features. Specifically, we performed a one-dimensional DCT in the time domain on the MFCC output over a moving window of *n* frames. This captured the modulation of each MFCC in the time domain in a different way than deltas do. Each DCT order (up to *n*) produced 20 new DCT-MFCC coefficients, which were appended to the original MFCCs. We kept the total number of dimensions to 40 (20 MFCCs + 20 delta-MFCCs). To achieve this, we

considered adding the zero-, first-, second-, or third-order DCTs to the original 20 MFCCs, with varying window sizes of 11, 21, and 31 frames (frames are computed every 10 ms). For the 100Ex condition, the best results were obtained by replacing deltas with third-order DCT coefficients with a 21-frame context window. For the 010Ex condition, the best results used the third-order DCT coefficients over an 11-frame window, instead of using deltas. In our submission, we used 21-frame DCT-based contextualization for both the 010Ex and 100Ex conditions to avoid relying on two different types of metadata.

Classification was done using a linear kernel SVM. Parameters were tuned using 10-fold cross-validation. The few videos with no audio track were considered to have a missing score for fusion purposes.

3.4 Audio Concept Recognition

The ACR subsystem is designed to detect auditory events, such as clapping, laughter, and hammering. The subsystem creates an acoustic event signature based on the output of 68 acoustic concept detectors run through the videos. This signature was used to run classification using an RBF-kernel SVM.

The 68 acoustic concept models were trained on annotations from SRI, the International Computer Science Institute, and Carnegie Mellon University that were done on the MED Research dataset. A linear SVM was used to train the audio concept classifiers using modified MFCCs with single deltas encoded to FVs using a 256-element GMM. The 68 classifiers were run every second on two-second intervals of the video to obtain "local responses" that indicated the prominence of each audio concept at different moments in the video. We also created a "global response" vector by running the 68 classifiers on the whole video. The metadata for the video consisted of these two responses.

We created a signature of 272 elements by concatenating the 68-dimensional global response with 3*68 local responses computed with three statistics for each concept: the average score across all frames, the average of the five top-scoring frames, and the minimum score of the five top-scoring frames.

An RBF-kernel SVM model was used for event classification. The SVM classifier was trained using 10-fold cross-validation to tune the regularization factor and to generate training scores for fusion. The few videos with no audio track were considered to have a missing score for fusion purposes.

3.5 Automatic Speech Recognition

The ASR subsystem uses an English ASR model trained on conversational telephone speech, and adapted to speech recorded in meetings. The basic audio segmentation and ASR capabilities are described in [8]. We performed supervised acoustic model adaptation to the TRECVID-MED domain using the LDC201208 release and unsupervised adaptation using first-pass recognition. We also performed supervised and unsupervised language-model adaptation to the TRECVID-MED domain. ASR was used to compute probabilistic word lattices from which we extracted video-based 1-gram word counts for MED. These counts were used to train a linear SVM with an L1 penalty. The resulting word counts, after stemming, formed the metadata.

The stemmed word counts were mapped to log counts using a soft cutoff of 1e-4. Counts from all the words were concatenated in a feature vector with a dimension of approximately 40,000. A linear SVM was then trained on these feature vectors using 10-fold cross-validation to tune the regularization factor and to generate training scores for fusion. The trained linear SVM was used to predict MED scores for videos with an expected word count above 1e-3. The videos that did not make this cutoff were considered to have a missing score for fusion purposes.

3.6 TextSearch

The TextSearch classifier uses the combined text detections from the 1-best output of both ASR and video OCR to determine whether a video is an event. SRI's video OCR software detects and recognizes text appearing in MED14 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 [9]. 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 Event Search (ES), we used a probabilistic information retrieval (PIR) approach, implemented with a Markov Random Field (MRF) model. In this method, the Event Kit is considered the query, and is scored against each test video clip. Conditional probabilities are obtained from frequency counts in English language Wikipedia text, with Laplacian noise modeling. Following Metzler and Croft (2005) [10], we applied Dirichlet smoothing to merge background statistics with those from the clip level. All input text was converted to lowercase and stemmed before use. Our MRF implementation permits dependencies between words to form phrases, and also permits certain terms or concepts in the query to be weighted more heavily than others.

The TextSearch semantic query generator uses a set of English language seed terms selected by a user from the Event Kit text description. We used circular similarities in a distributed word space (computed with word2vec [11] from Wikipedia) to identify additional terms that are similar or related to the seed terms. To facilitate rapid selection and removal of terms, we further clustered the additional terms to identify thematic groups. Each group was then described by 1-3 centroids to provide a quick summary of the contents. Figure 2 provides an example of the thematic grouping of terms for Event E021.

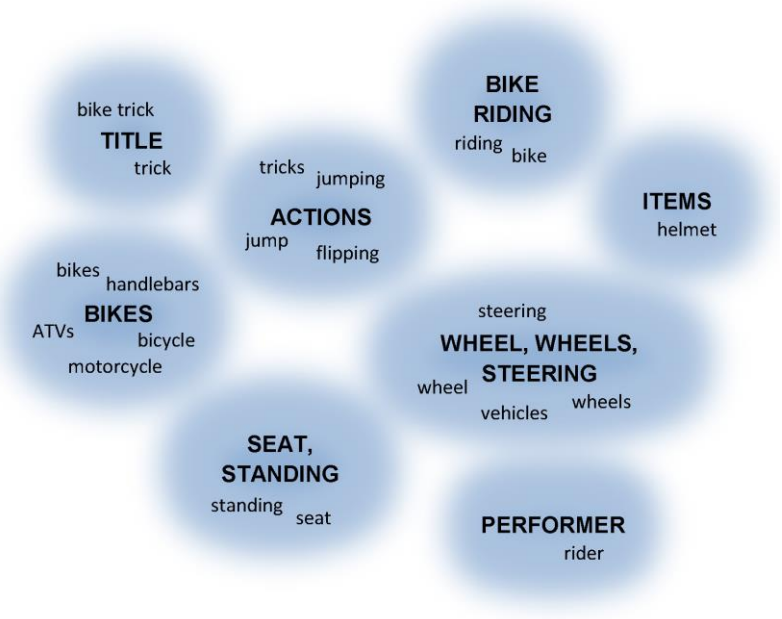

Figure 2. Thematic grouping of semantic query terms for Event E021, "Attempting a Bike Trick".

3.7 Fusion and Thresholding

We used the LLR late fusion approach [12], 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 whether a video is a positive instance of the specified event. The LLR fusion approach computes different fusion weights for each classifier and for each event. The training data for the fusion weights were the detection scores for each classifier, obtained by running 10-fold cross-validation with the Event Kit positives and the Event Background negatives. This procedure provided an unbiased detection score for each trial in the 010Ex 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 classifiers. We included an indicator variable for each classifier to account for the possibility of missing scores in some trials.

We then used logistic regression to train fusion models. The zero- and first-order statistics of scores for each classifier and each event were computed on the cross-validation data and used to normalize the scores to be zero-mean and unit-variance during both fusion training and ES. Normalizing ensures that the scores are comparable across classifiers, which makes logistic regression work better. During training, if the weight of a classifier was found to be negative, logistic regression was retrained with that classifier removed. The training log-odds were computed from the priors and were stored with the model to be used at search time to obtain the LLRs.

During ES for 100Ex and 010Ex, the detection scores from every classifier were concatenated into a feature vector for each test video, similar to what was done during training. Binary indicators for missing scores were appended. Applying the fusion model to this feature vector gave a log-posterior, which was converted to an LLR by subtracting the training log-odds. For 010Ex, the selected threshold was the optimal Bayesian threshold for R_0, assuming a similar prior on the training and test data. For 100Ex, we selected the threshold that maximized R_0 on the training data. These two different choices for 010Ex and 100Ex were motivated by observed performance on the test set, but the techniques provided comparable performance.

During ES for SQ, the event detection scores from two classifiers (VideoStory and TextSearch) were Z-normalized using statistics computed on the MED Research dataset on events E001-5, then mapped to [0,1] using a sigmoid function. Mean score averaging gave the final fused score, with a score of 0 assumed for missing scores. Thresholds were computed using the MED Research dataset on E001-5.

3.8 MED Evaluation

Table 1 shows the MED performance of the SESAME system (the metric is mean average precision) on the MED14-MEDEvalFull and MED14-EvalSub datasets. Table 2 shows the MED performance in terms of R0, the minimum acceptable recall, on the MED14-MEDEvalFull and MED14-EvalSub datasets.

Table 1: SESAME System MED Performance (Mean Average Precision) on MED14-MEDEvalFull and MED14-EvalSub Datasets

	MED14Eval Full		MED14Eval Sub	
	PS	AH	PS	AH
100Ex	29.9%	32.8%*	38.1%	40.6%*
010Ex	18.3%	16.9%*	23.7%	24.1%*
SQ	5.1%	2.4%	8.6%	4.9%

** Debugged submission*

Table 2: SESAME System MED Performance (in Terms of R0, the Minimum Acceptable Recall) on MED14-MEDEvalFull and MED14-EvalSub Datasets

	MED14Eval Full		MED14Eval Sub	
	PS	AH	PS	AH
100Ex	58.7%	56.2%*	57.0%	50.7%*
010Ex	41.9%	34.8%*	39.7%	30.8%*
SQ	< 0	2.9%	< 0	1.4%

** Debugged submission*

Table 3 shows the MED performance of the SESAME system by individual classifier on the MED14-Test dataset. The fusion weights used to generate these results are the same ones used to generate Table 1 and Table 2. We discovered, however, that some detected videos were falsely labeled as negatives. Table 4 shows the MED performance by individual classifier on the same dataset after correcting the ground truth annotations. We drew these conclusions:

- Low- and high-level visual features, low-level motion features, and VideoStory contributed most significantly to event detection performance.
- For 100Ex and 010Ex, the LLR late-fusion method significantly improved performance over that of the best individual classifier.

- For SQ, fusion is worse than TextSearch in Table 3. This emphasizes the difficulty of using fusion modalities without training data to either normalize or weight the scores. The fusion parameters were tuned on the MED Research dataset, but those parameters don't appear to be optimal for the MED14-Test dataset. Therefore, when we reran the experiment after correcting the ground truth annotations, for SQ, we used equal fusion weights because of the absence of adequate training data. This strategy resulted in fusion results that were better than results for both TextSearch and VideoStory.

Table 3: SESAME System MED Performance (Mean Average Precision) by Classifier on the MED14-Test Dataset for E021 to E030

Condition	100Ex	010Ex	SQ
ACR	4.0%	1.5%	
ASR	5.1%	0.9%	
MFCC	8.6%	3.5%	
DTFV	23.6%	10.7%	
Visual High	19.8%	12.5%	
Visual Low	23.2%	11.9%	
VideoStory	25.8%	15.1%	2.3%
TextSearch	3.9%	3.9%	3.9%
FUSION	**36.2%**	**22.4%**	**3.5%**

Table 4: SESAME System MED Performance by Individual Classifier on the Table 3 Dataset with Corrected Ground Truth Annotations

Condition	100Ex	010Ex	SQ
ACR	4.0%	1.4%	
ASR	5.1%	0.9%	
MFCC	10.3%	3.9%	
DTFV	32.0%	14.7%	
Visual High	29.0%	16.6%	
Visual Low	32.0%	15.8%	
VideoStory	35.5%	20.7%	6.2%
TextSearch	3.9%	3.9%	3.9%
FUSION	**51.9%**	**27.5%**	**7.8%**

4. MULTIMEDIA EVENT RECOUNTING

4.1 MER Description

The goal of MER is to provide a human-understandable recounting for each positive clip that the MED system detected. The recounting evidence consists of concepts in the SQ that were found and identified as key evidence in the search video, indicating that the video contains the event. The recounting includes a confidence value and localization information (time interval and spatial location) for each piece of key evidence.

The SESAME system generated event recountings for the 010Ex condition based on semantic concepts from TextSearch and the sets of automatically detected concepts (15,000 visual concepts and 164 action concepts.) VideoStory was not included in the MER process because no mechanism was available at the time to generate recounting information for its individual concepts.

For the TextSearch dataset, key evidence included the highest-scoring terms in the PIR model for the event, and the confidence score was the probability of the term given the event model. The highest-scoring visual and action concepts that had been selected for the SQ were also selected as key evidence. However, the highest-scoring concepts often occurred at lower levels in the concept hierarchy. For example, if the concept "dog" were in the SQ, we would want to increase its confidence score if the detectors for one or more dog breeds produced a significant response. We computed the score of these concepts by aggregating scores from the lower levels in the concept tree.

Although the MED performance of the 15,000 visual concepts in aggregate is quite good, the reliability of detectors for most of the individual visual and action concepts, and for determining the time interval in the video where a concept appears, is not very good. Therefore, we applied a strategy to augment the response of an individual concept detector by using the entire MED response to choose the context within the video where the relevant concepts are more likely to be found. To find intervals with the most salient context, our ELM approach produced the best results, but it was not computationally feasible to run it on a COTS workstation for

event query generation and ES during the TRECVID MED and MER evaluations. Therefore, we decided to use an SVM with a histogram intersection kernel. The classifier was trained separately on action and object concepts at the video level. Then, the learned classifiers were applied to short segments, and the highest-scoring segments were selected for recounting. The highest-scoring concepts within these intervals were presented as MER evidence.

4.2 MER Evaluation

The purpose of the MER evaluation was to assess the quality of recounting evidence associated with the MED retrieval results. The MER evaluation was conducted with the help of human judges. The event query and recountings were assessed according to (1) whether the semantic query seemed like a concise and logical query that would be generated for the event description; (2) how well the key evidence convinced the judge of the occurrence of the event in the video; and (3) how compact the key evidence snippets were, compared with the length of the video. Table 5 shows the evaluation of the SESAME system according to the first two criteria, which were judged on a Likert-style scale. The ratio of the duration of key evidence snippets to the length of the original video was 19.7%.

Table 5: MER Performance of the SESAME System (Likert Scale)

Criteria	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
Query Conciseness	6%	10%	12%	54%	17%
Key Evidence Convincing	24%	13%	10%	30%	22%

5. ACKNOWLEDGEMENTS

This work was supported by the Intelligence Advanced Research Projects Activity (IARPA) via the Department of Interior National Business Center (DoI/NBC), contract number D11PC20067. 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, M. Michel, J. Fiscus, G. Sanders, B. Shaw, W. Kraaij, A.F. Smeaton, and G. Quéenot, TRECVID 2014 – An Overview of the Goals, Tasks, Data, Evaluation Mechanisms, and Metrics, *Proceedings of TRECVID 2014*, http://www-nlpir.nist.gov/projects/tvpubs/tv14.papers/tv14overview.pdf

[2] Habibian, A., T. Mensink, and C.G.M. Snoek, VideoStory: A New Multimedia Embedding for Few-Example Recognition and Translation of Events, *ACM Multimedia*, 2014.

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

[4] Center for Research in Computer Vision, UCF101 – Action Recognition Data Set, http://crcv.ucf.edu/data/UCF101.php, 2013.

[5] Soomro, K., R.A. Zamir, and M. 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.

[6] Sadanand, S. and J. J. Corso, Action bank: A high-level representation of activity in video, *2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[7] Sun, C. and R. Nevatia, DISCOVER: Discovering Important Segments for Classification of Video Events and Recounting, *2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[8] van Hout, J., M. Akbacak, D. Castan, E. Yeh, and M. Sanchez, Extracting spoken and acoustic concepts for multimedia event detection, *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2013.

[9] 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), pp. 147-158, July 2005.

[10] Metzler, D. and W.B. Croft, A Markov Random Field Model for Term Dependencies, *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2005)*, pp. 472-479, 2005.

[11] word2vec, https://code.google.com/p/word2vec/.

[12] van Hout, J., E. Yeh, D.C. Koelma, C.G.M. Snoek, C. Sun, R. Nevatia, J. Wong, and G.K. Myers, Late Fusion and Calibration for Multimedia Event Detection Using Few Examples, *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2014.