The 2012 SESAME Multimedia Event Detection (MED) and Multimedia Event Recounting (MER) Systems

SRI International (SRI)
Murat Akbacak, Robert C. Bolles, J. Brian Burns, Mark Eliot, Aaron Heller, James A. Herson, Gregory K. Myers, Ramesh Nallapati, Stephanie Pancoast, Julien van Hout, Eric Yeh

University of Amsterdam (UvA)
Amirhossein Habibian, Dennis C. Koelma, Zhenyang Li, Masoud Mazloom, Silvia-Laura Pintea, Koen E.A. van de Sande, 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 four runs for the MED12 pre-specified events, two runs for the ad hoc events, and a run for multimedia event recounting. The detection runs included combinations of low-level visual, motion, and audio features; high-level semantic visual concepts; and text-based modalities (automatic speech recognition [ASR] and video optical character recognition [OCR]). The individual types of features and concepts produced a total of 14 event classifiers. We combined the event detection results for these classifiers using three fusion methods, two of which relied on the particular set of detection scores that were available for each video clip. In addition, we applied three methods for selecting the detection threshold. Performance on the ad hoc events was comparable to that for the pre-specified events. Low-level visual features were the strongest performers across all training conditions and events. However, detectors based on visual concepts and low-level, motion-based features were very competitive in performance. Table 1 summarizes the runs:

Table 1: Pre-specified and Ad Hoc Runs for MED12

Run	Event Set	Training Set	Event Classifiers	Fusion Method	Thresholding
1	Pre-specified	EkFull	All exc. VOCR and ASR event kit text	WMR	score@TER
2	Pre-specified	EkFull	All exc. VOCR and ASR	SparseEM	score@TER
3	Pre-specified	EkFull	All	MAP	Box_avg
4	Pre-specified	EK10Ex	All exc. VOCR and ASR	MAP	score@TER
5	Ad hoc	EkFull	All exc. VOCR and ASR	MAP	score@TER
6	Ad hoc	EK10Ex	All exc. VOCR and ASR	MAP	median score@TER

1. INTRODUCTION

The purpose of the MED12 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 contents of the videos in these collections are 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.

Each event was defined by an *event kit*, which consisted of an event name, definition, explication (textual exposition of the terms and concepts), evidential descriptions, and illustrative video exemplars. Systems in the MED evaluation used the event kit to generate an *event agent*, which identified videos in the target video collection that contained the specified event. The event agent was executed on metadata that had been pre-computed from the video collection. The goal of a MED system was to achieve event detection performance with low miss and false-alarm rates, and to achieve an operating point that corresponds to a ratio of miss and false-alarm rates near a specified target error ratio (TER).

The MED evaluation consisted of two types of events: pre-specified events and ad hoc events. For pre-specified events, the metadata could be extracted from the videos with knowledge of the pre-specified event kits. For ad hoc events, the metadata was extracted from the videos *without* knowledge of the ad hoc test event kits, and event detection had to be completed in a much shorter amount of time. Performance on ad hoc events is an indication of the robustness of the system for handling subsequent user-defined events, when metadata generation cannot be optimized for a known set of events.

The number of video exemplars used for training event agents affects system performance. MED12 specified the use of two levels of exemplars: EKFull, which used the actual event definition and all provided videos; and EK10Ex, which used a specified set of ten positive and ten related video exemplars. (In addition, other videos that were not positives or related to other events could be used as negatives for training.)

To handle this challenging problem, the SESAME system extracted a comprehensive set of heterogeneous low-level visual, audio, and motion features. SESAME also extracted higher-level semantic content in the form of visual concepts, spoken text, and video text. As the SESAME system matures, it will include additional higher-level, semantic-content extraction capabilities.

The SESAME event agents included a total of 14 event classifiers, each using one type of feature or semantic content:

- Two event classifiers were based on low-level visual features.
- Five event classifiers used low-level motion features.
- One event classifier was based on low-level audio features.
- Two event classifiers were based on visual concepts.
- Four event classifiers used the text results from ASR and video OCR processes.

We combined the event detection results for these classifiers using three fusion methods, two of which relied on the particular set of detection scores that were available for each video clip. In addition, we applied three methods of selecting the detection threshold.

This paper describes the content extraction methods (Section 2), fusion methods for the runs (Section 3), threshold selection methods (Section 4), and the experimental results (Section 5).

2. CONTENT EXTRACTION METHODS

2.1 Visual Features

Two event classifiers were based on low-level visual features. The visual features consisted of vector-quantized Harris-Laplace and dense-sampled color SIFT descriptors [2] computed on frames sampled once every two seconds. All descriptors were reduced by principal component analysis (PCA). We employed spatial pyramids and average or difference coding [3]. We aggregated code word histograms per frame to video level. For learning event models, we used either a non-linear support vector machine (SVM) with a fast histogram intersection kernel [4] or a linear SVM [5].

2.2 Motion Features

Five event classifiers used low-level motion features based on STIP [6], dense trajectories (DTs) [7], and MoSIFT [8]. We computed the STIP features and DT features at two spatial resolutions and then encoded using first- and second-order Fisher Vector descriptors [3]. Descriptors were aggregated across each video. Four event classifiers were generated: two with the STIP features using first-order and second-order Fisher Vector descriptors, and two with the DT features using first-order and second-order Fisher Vector descriptors. The event classifiers were trained on the SESAME challenge training set by five-fold cross-validation and were tested on a previously defined validation set. An SVM with a Gaussian kernel performed the classification for the MED12 task. A fifth classifier, based on MoSIFT features [3] provided by CMU, was trained on the same training set using an SVM with a $\chi 2$ kernel.

2.3 Audio Features

The audio data was extracted from the video files and converted to a 16 kHz sampling rate. Mel frequency cepstral coefficients (MFCCs) were computed at 10 ms intervals. A 39-dimensional feature vector, consisting of these 13 values (12 coefficients and the log-energy) along with their delta and delta-delta values, was computed. These features' vectors were then vector-quantized into code words with a 1,000-word code book (which was determined from the MED11 development set). A histogram of the code words was generated over the entire clip. An SVM classifier with a histogram intersection kernel was trained on the video exemplars from the event kits and used to detect the events.

2.4 Visual Concepts

Two event classifiers were based on concept detectors. Our 2012 event representation contains 346 concepts based on training data from the TRECVID 2012 SIN task, and 1,000 concepts based on training data from the Pascal ImageNet collection. The concepts vary from scenes like *landscape*, objects like *motorcycle*, actions like *throwing*, and people like *male*. All detectors were trained using Fisher difference coding of color SIFT with a linear SVM. One event classifier used random forests, and the other used a non-linear SVM [5].

2.5 Automatic Speech Recognition (ASR)

We used SRI's ASR software to recognize spoken English. From the recognized speech, we generated two event classifiers: one employed a simple bag-of-words model against the combined word content of the clip, and the other matched the recognized speech with a set of keywords specific to the event. For both of these classifiers, the text of the recognized speech was converted to lowercase, stemming was applied, and stop words were removed. For the keyword-based event classifier, we used an inverse document frequency (IDF) measure to identify the top 20 most relevant terms from the event explication, and for each event, we added terms that were not in the kit. These additional terms were deemed likely to appear in videos, and were developed independently of the clips and clip metadata. The classifiers were trained using logistic regression.

2.6 Video OCR

SRI's video OCR software recognized and extracted text from MED-12 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 detects and reads printed text and was configured to recognize English language text. We generated two event classifiers similar to ASR: one employed a simple bag-of-words model against the combined word content of the clip, and the other matched the recognized text with a set of keywords specific to the event.

3. FUSION METHODS

We tested several late fusion methods in our experiments on our internal validation set. For all our fusion experiments on the validation set, we trained the event classifiers corresponding to each modality on our internal training set, and executed the classifier on the validation set to produce detection scores for each event. Then, we produced detection scores on the validation set using the fusion model with 10-fold cross validation on a per-event basis. The fusion models we tested are summarized below.

- **Conditional mixture model:** This model combined the detection scores from various modalities using mixture weights that were trained by the expectation maximization (EM) algorithm on the labeled training folds. For clips that were missing scores from one or more modalities, we filled in the missing scores with the expected score from that modality based on the training data.
- **Sparse conditional mixture model:** This was an extension of the conditional mixture model that addressed the problem of missing scores for a clip by computing a mixture for only the observed modalities [10]. This was done by renormalizing the mixture weights over the observed modalities for each clip. The training was done with the EM algorithm; however, the maximization step no longer had a closed form solution, so we used gradient descent techniques to learn the optimal weights.
- **Geometric mean (GM):** In this method, we computed the uniform geometric mean of the scores of the observed modalities for a given clip.
- **SVMLight:** This fusion model consisted of training an SVM using the scores from various modalities as the features for each clip. We used the SVMLight[1] implementation with linear kernels.
- **LibSVM_quad:** This fusion model was same as SVMLight, but we used the LibSVM[2] toolkit with polynomial kernels of degree 2.

[1] http://svmlight.joachims.org/

[2] http://www.csie.ntu.edu.tw/~cjlin/libsvm/

- **BBN_weighting (BBN):** This was a weighted averaging method described in [11] that dynamically adjusted the weights of each modality for each clip based on how far the modality's score was from its decision threshold.
- **Mean average precision-weighted fusion (MAP):** This was another weighted fusion method that weighed scores from the observed modalities for a clip by their normalized average precision scores, as computed on the training fold.
- **Weighted mean root:** The weights were average precision scores using 10-fold cross-validation on the validation data. For each video (x), we performed a power normalization using:

  $$\boldsymbol{x}^{\frac{1}{\alpha}} \tag{1}$$

  where α was determined by the number of event scores for that video.

The following subsections summarize our fusion experiments on our internal validation set for each of the four training conditions.

3.1 Pre-specified Events, EKFull Condition

The event classifiers for each modality were trained on all pre-specified events using the *EKFull* condition. The visual SIFT-based modalities (both bag-of-words and difference coding) are our best performers, followed by the visual-concepts-based modalities and the motion-based modalities (dense-trajectory, MOSIFT, and STIP-Fisher), and the audio-based MFCC modality. Because VOCR- and ASR-based scores were very sparse (missing scores for many clips), they did not carry a strong signal for event detection. We excluded the VOCR and ASR modalities from the fusion model because they had very weak signals and did not contribute much to the fusion.

Our experiments indicated that the MAP weighting, sparse conditional mixture model, and geometric mean are the best performing fusion models. All these models achieved an improvement of about 33% compared to the best single-modality run in terms of MAP. Our experiments on other training conditions (not reported here) indicated that MAP weighting was the most stable of the fusion models. We therefore chose MAP weighting as the default fusion model in the rest of our experiments.

3.2 Pre-specified Events, EK10Ex Training Condition

For the EK10Ex training condition for pre-specified events, we trained our event classifiers corresponding to each modality on only 10 positive clips in our training set, and generated detection scores on the validation set. We evaluated the performance of the MAP-based fusion model on the validation set using 10-fold cross validation.

The performance of each modality was considerably lower, due to a smaller amount of training data available to the event classifiers, but the relative importance of various modalities remained about the same as for the EKFull run. However, in terms of relative improvement in performance, the fusion model was nearly 75% better than the best individual modality.

3.3 Ad Hoc Events, EKFull Training Condition

For the ad hoc events, we created our own internal partition of training and validation sets by splitting the ad hoc events kit event-wise, and adding positive clips from pre-specified events as negative clips for the ad-hoc events on both training and validation sets. We repeated our fusion experiments using 10-fold cross validation on the new validation set as described for the pre-specified setting.

The performance numbers for the fusion model and its improvement over the best performing single modality were comparable to the EKFull runs for pre-specified events. However, the relative importance of the modalities had changed to some extent. For the ad hoc events, the visual-concepts-based detectors played a more prominent role than for the pre-specified events, and were better performers than the motion-based detectors.

3.4 Ad Hoc Events, EK10Ex Training Condition

Similar to the EK10Ex condition for the pre-specified events, we trained our event classifiers on a subset of 10 positive clips for each of the ad hoc events and generated the detection scores on our validation set. We compared the performance of the fusion model with that of each of the individual modalities on the validation set.

We noticed that the visual-concepts-based modalities were strong performers, similar to the EKFull training condition for the ad hoc events. However, the relative improvement in performance of the fusion model was not as significant in this case compared to the EK10Ex training condition for the pre-specified events.

When we ran the fusion model on the Progress Set for the official submission, we did not use the fusion model that was trained on the ad hoc events, because it included training on additional positive clips from the validation set, which was a violation of the Ek10Ex requirement. We trained the fusion model on the Ek10Ex condition on the pre-specified events, averaged the learned weights across all pre-specified events, and used them as the common set of fusion weights for all ad hoc events on the Progress Set.

3.5 Event-specific DET Curves

Figure 1 shows event-specific DET curves of our MAP-based fusion model computed on an internal validation set for the four conditions described previously. The results were very similar: DET curves for the EK10Ex condition were poorer than for the EKFull condition for both pre-specified and ad hoc events. In addition, the performance on the ad hoc events was as robust as that on the pre-specified events. This is not surprising, because our method of training event agents was identical for both types of events, and adequate time was allowed for training and execution for the ad hoc events.

4. AUTOMATIC THRESHOLD SELECTION

As per the MED guidelines, the performers were allowed to tune only the threshold before running the event agent on the progress set. This meant that the threshold-selection algorithm could not be based in any way on the distribution of scores on the Progress Set. Therefore, we used the score at TER (score@TER), as computed on a per-event basis, as the threshold-selection algorithm on our validation set.

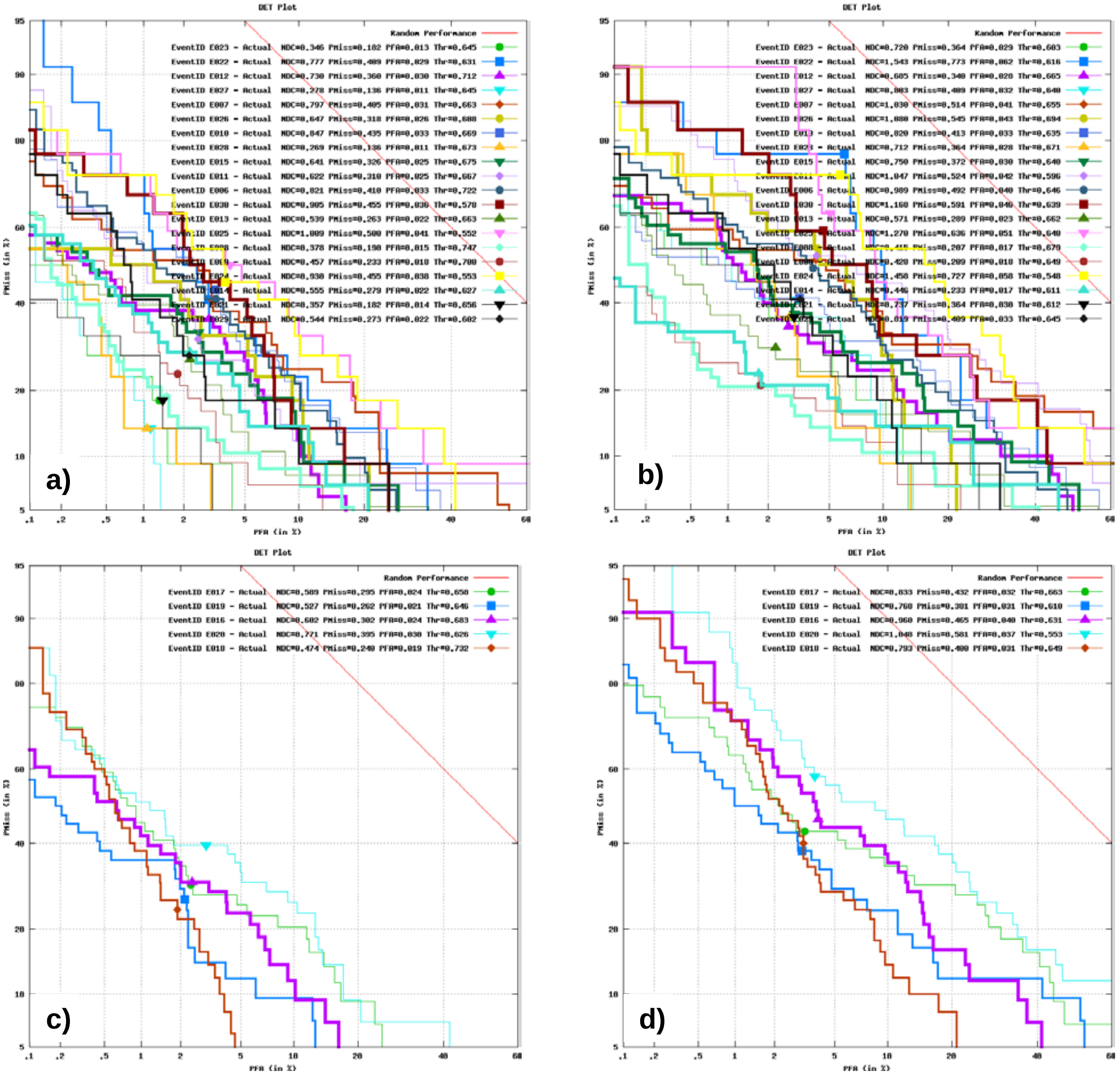

Figure 1. Event-specific DET curves for the MAP-based fusion model on a validation set for (a) pre-specified events and EKFull condition, (b) pre-specified events and EK10Ex condition, (c) ad hoc events and EKFull condition, and (d) ad hoc events and EK10Ex condition.

We used a second method of computing the threshold to try to ensure that the operating point would fall within the bounds of particular miss and false-alarm probabilities. The box-averaging threshold was computed based on the average of the following two threshold estimates based on our internal validation set: one threshold was computed at the intersection of the DET curve with the horizontal pMD=50% line; the other threshold was computed at the intersection of the DET curve with the vertical pFA=4% line.

For the ad hoc Ek10Ex condition, we were not allowed to train either the fusion model or the threshold selection algorithm on additional positive clips from the ad hoc events. Therefore, we used a single threshold for all ad hoc events. This threshold was the median of the score@TER thresholds we learned on the pre-specified events for the EK10Ex condition.

5. EXPERIMENTAL RESULTS AND CONCLUSIONS FOR MED

Table 2 shows the official evaluation results for our six submissions.

Run1 used the weighted mean-root-fusion method with all the modalities except VOCR and ASR, and Score@TER for threshold selection, and performed the best on pre-specified events on the EKFull training condition. Run2, which used the sparse mixture model with the same modalities and threshold selection method, did not perform as well. The performance of Run3, which used the MAP-weighted fusion of all modalities including ASR and VOCR, and box-averaging for threshold selection, was comparable to the others.

Table 2: SESAME MED 12 Performance

Run	P(FA)	P(Miss)
1	2.94%	22.35%
2	4.89%	16.07%
3	4.05%	19.53%
4	6.94%	36.03%
5	3.62%	20.04%
6	1.97%	56.98%

Not surprisingly, the performance of Run4 on the pre-specified EK10Ex condition was lower than that for EKFull, since the event agents received less training data. The performance of Run5, on the ad hoc EKFull condition, was comparable to that of the EKFull condition for pre-specified events.

The average miss rate of Run6 for the ad hoc events on the EK10Ex condition was considerably higher than in other runs. We suspect this was a result of tuning the thresholds on the pre-specified events and using the median of those thresholds as the predicted threshold for all ad-hoc events, which did not transfer very well.

In terms of performance of various modalities, low-level visual features were the strongest performers across all training conditions and events. However, we found that this year, our visual-concepts-based detectors and low-level motion-based detectors were very competitive in performance. Low-level audio features provided a weak but useful signal for event detection. We still need to improve our ASR- and VOCR-based modalities.

Also, contrary to our initial perception, we found that simple MAP-based fusion was very effective in late fusion, compared to more sophisticated discriminative or generative models.

6. MULTIMEDIA EVENT RECOUNTING

For the multimedia event recounting task, we extracted data related to semantic concepts: ASR, video OCR, camera motion statistics, face statistics, and visual concepts. We first describe the recounting methods using ASR, video OCR, camera motion, and face data, and then provide details for the recounting procedures with visual concepts. Finally, we describe the results of two recounting experiments and the TRECVID MER evaluation.

6.1 Recounting Using ASR, Video OCR, Camera Motion, and Face Data

ASR: We created a list of specific words for each event class by combining the text found in the event kits and the text associated with each positive video clip in the training set. For video clips that had speech, we extracted the ASR words and their starting time and duration. To reduce the amount of recountable text, we used a five-step procedure to remove unnecessary words:

1) Removed stop words.
2) Eliminated words with a length of two characters or less, as these are either erroneous or non-informative.
3) Eliminated repetitive words.

4) Removed words that consisted of invalid characters, or words in which the number of non-alphabet characters was more than 0.2 of the length of the words.
5) If the video clip had an event label, we removed the words that were not in the list of specific words for this event. If the video clip didn't have an event label, we removed the words that were not in the list of specific words for all events.

After the final step, we had a list of informative ASR words to report in the output of our event recounting.

VOCR: As for ASR, we created a list of specific words for each event class by combining the text found in the event kits and the text associated with each positive video clip in the training set. For video clips with overlay text identified by SRI's video OCR, we extracted the text words, the starting and ending frame indices of the text words, and their confidence scores. We reduced the amount of recountable text using the approach followed for ASR. We also removed words that our VOCR detected with a low confidence value. After these steps, we prepared a list of informative VOCR words to report in the output of our event recounting.

Camera motion: For video clips with camera motion, as indicated by SRI's camera motion detector, we extracted the type of camera motions, their starting and ending frame indices, and their confidence scores. When camera motion was detected, we reported these extracted data in the output of event recounting.

Faces: For video clips in which faces were detected, we reported the following extracted data for event recounting: the total number of faces detected, the number of faces larger than 1/3 the height of the frame, the average detection confidence for big faces, the number of faces that were smaller than 1/3 the height of the frame, the average detection confidence for small faces, and the fraction of video clips with no faces.

6.2 Recounting with Visual Concepts

For visual concepts, we selected the concepts of interest based on our pool of 1,346 concepts as defined in our CDR. We manually assigned each concept to one of the following recounting categories: objects, actions, scene, people, animals, and others. Out of 1,346 concepts, we found 705 objects, 26 actions, 135 scenes, 82 people, 338 animals, and 60 concepts for the others category. We focused on the objects, actions, scenes, and people recounting categories. We used two strategies to select concepts of interest: one based on visual selection using the provided positive videos, and one based on textual selection using the provided event kit descriptions.

Recounting using example video. We extracted one frame per two seconds for each video. We applied all 1,346 concept detectors from our CDR to the frames, resulting in a vector of 1,346 concept-detector probabilities. We averaged the normalized scores over all frames per video to obtain a unique semantic representation for each video. To identify the informative concepts as key observations in each video clip for each event category, we used logistic regression with L1 regularization. L1-regularized logistic regression involves the following unconstrained optimization problem:

$$\min_w f(w) = ||w||_1 + \ C \textstyle\sum_{i=1}^{l} \log(1 + e^{-y_i w^T x_i}) \tag{2}$$

where $(x_i, y_i), i = 1, \ldots, l$ is a set of instance-label pairs, $\log(1 + e^{-y_i w^T x_i})$ is a non-negative (convex) loss function, $||w||_1$ is a regularization term that avoids over-fitting the training data, and $C > 0$ is a critical user-defined parameter that balances the regularization and loss terms.

The output of this L1-regularized formula is a sparse vector w, with relatively few nonzero coefficients. The nonzero elements of *w* helped us identify the important concepts. When $w_j = 0$, it meant that the logistic model did not use the *j*th component of the feature vector, so in our case, a sparse *w* corresponded to a logistic model that depended on only a few concepts.

- To select the best parameter *C*, corresponding to the best set of selected concepts for each event category, we conducted five-fold cross validations varying *C* from 1 to 200.
- For validating the selected concepts, we trained a linear support vector classifier on the training set and evaluated this classifier on the validation set.
- For evaluating the model, we considered MAP as our evaluation criteria.

After finding the best set of concepts for each event class, we sorted them for each of the recounting categories, and identified which objects, actions, scenes, and people concepts were important to consider for the recounting.

Recounting using event kit description. For each event class, we had a textual event kit that included information about the event, such as event name; definition; evidential description that included which objects, people, activities, and scenes could be expected to occur in the positive video clips for this event class; and information about audio. We mapped the event kit text of each event class into our 1,346 semantic concepts. After removing stop words from the event kit, we computed the similarity distances between all concepts and words using WordNet. To be precise, the inputs of our method were the concepts and the event kit for each event.

The output of our method was the similarity distance of each concept by each event. For each concept, we obtained a vector of distances between the concept and all the words in the event kit. We considered the minimum value of this vector to be the similarity distance between the concept and the event. We repeated this procedure for all concepts and all event kits. After converting the event kit of each event class to an L1-normalized vector of real numbers, we sorted this vector. Since each concept had a recounting category, there was a sorted list of objects, actions, scenes, and people concepts for each event. We selected concepts from the sorted list based on how many objects, actions, scenes, and people concepts we wanted to report in the recounting.

6.3 MER Experiments

To assess the effectiveness of these recounting methods, we ran two experiments similar to the TRECVID MER evaluation tasks. Figure 2a shows the multimedia event recounting results for the task in which human judges assigned the appropriate event category to a recounting of a video clip. We observed that, for this task, a recounting consisting of concepts selected using text from the event kits appeared to be more informative than a recounting based on selecting concepts from example videos. On average, the textual method was 49% correct, and the visual method 45% correct. In four out of five events, the textual method was better than the visual method, especially for the event *renovating a home*, where the absolute difference was 22%. For the other events, the difference was less extreme. For the event *working on a metal crafts project*, the visual method was more appropriate. We concluded that the judges, including analysts, preferred an event recounting close to the provided textual event kit, but the difference from visual selection was not that great on average.

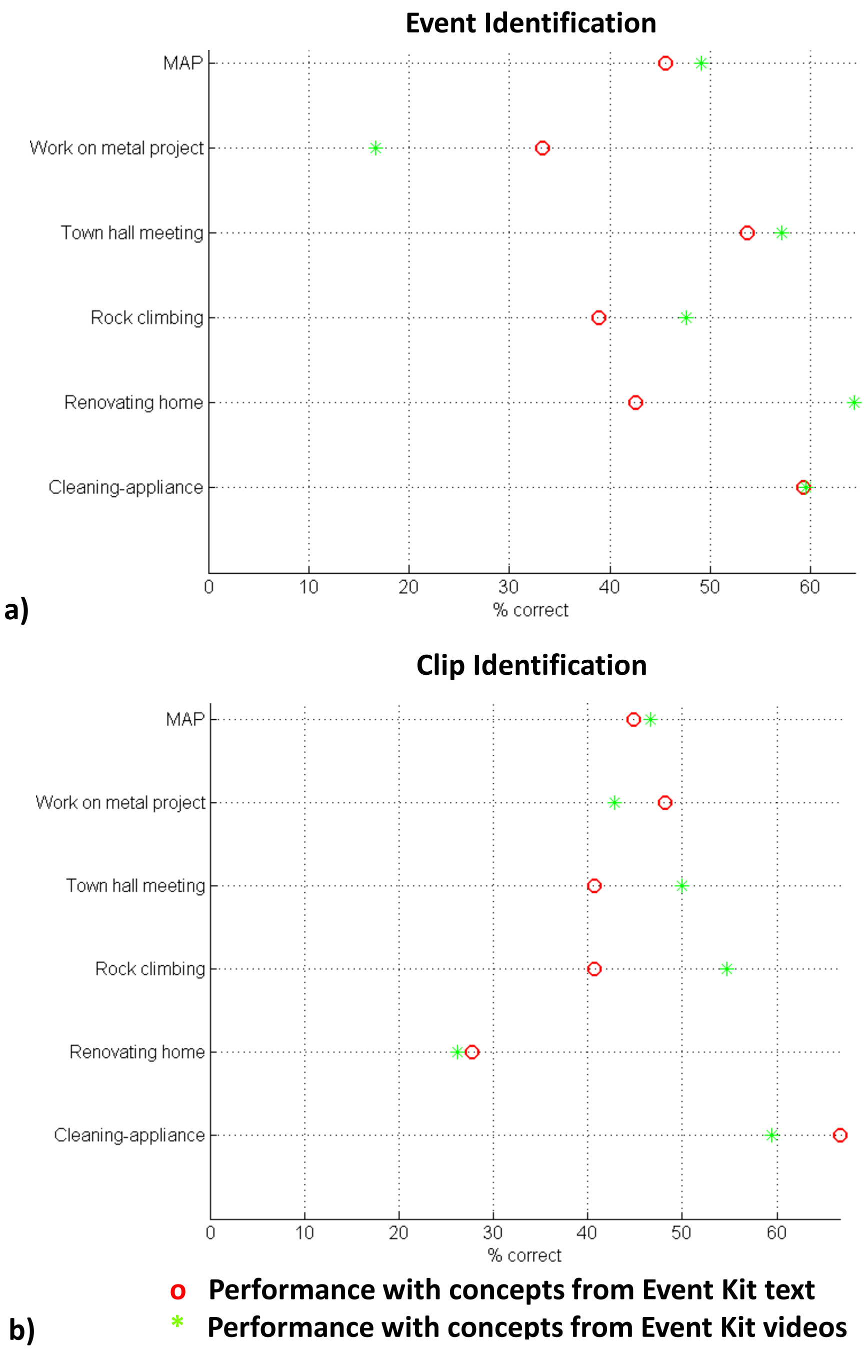

Figure 2. User experiment to evaluate MER performance.

Figure 2b shows the multimedia event recounting results for the task in which judges assigned an event recounting of a video to the appropriate video clip. We observed that for this task as well, a recounting consisting of concepts selected using text from the event kits was slightly more informative than a recounting based on selecting concepts from event example videos. On average, the textual method was 47% correct, and the visual method 45% correct. However, in three out of five events, the visual method was better than the textual method. We concluded that the judges had different preferences: for some events, the visual method was preferred, and for others, textual selection was preferred.

Similar results were obtained in the TRECVID MER evaluation tasks. For event identification, 45.56% of the judgments correctly matched the MER output to the event kit. For clip identification, 44.81% of the judgments correctly matched the MER output to the clip.

7. ACKNOWLEDGEMENT

This work was supported by the Intelligence Advanced Research Projects Activity (IARPA) via the Department of Interior National Business Center, 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.

8. REFERENCES

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

[2] van de Sande, K.E.A., Gevers, T., Snoek, C.G.M. "Evaluating Color Descriptors for Object and Scene Recognition. " *IEEE Trans. Pattern Analysis and Machine Intelligence*, 32(2), 2010.

[3] Jégou, H., Perronnin, F., Douze, M., Sanchez, J., Pérez, P., Schmid, C. "Aggregating Local Image Descriptors into Compact Codes." *IEEE Trans. Pattern Analysis and Machine Intelligence*, 34(9), 2012.

[4] Maji, S., Berg, A.C., Malik, J. "Efficient Classification for Additive Kernel SVMs." *IEEE Trans. Pattern Analysis and Machine Intelligence*, in press.

[5] Snoek, C.G.M. et al. "The MediaMill TRECVID 2012 Semantic Video Search Engine." *Proc. TRECVID Workshop*, Gaithersburg, MD, 2012.

[6] Laptev, I. "On Space-Time Interest Points." *International Journal of Computer Vision*, 64(2/3), pp.107-123, 2005.

[7] Wang, H., Kläser, A., Schmid C., Cheng-Lin, L. "Action Recognition by Dense Trajectories." *CVPR*, 2011.

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

[9] Myers, G., Bolles, R., Luong, Q.-T., Herson, J., Aradhye, H. "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] Nallapati, R., Yeh, E., Myers, G. "Sparse Mixture Model: Late Fusion with Missing Scores for Multimedia Event Detection." *SPIE Multimedia Content Access: Algorithms and Systems VII*, 2013.

[11] Natarajan, P., Wu, S., Vitaladevuni, S., Zhuang, X., Tsakalidis, S., Park, U., Prasad, R. "Multimodal Feature Fusion for Robust Event Detection in Web Videos." *CVPR*, 2012.