TRECVID 2013 GENIE:
Multimedia Event Detection and Recounting

Sangmin Oh[1], A. G. Amitha Perera[1], Ilseo Kim[1], Megha Pandey[1],
Kevin Cannons[2], Hossein Hajimirsadeghi[2], Arash Vahdat[2], Greg Mori[2],
Ben Miller[3], Scott McCloskey[3], You-Chi Cheng[4], Zhen Huang[4], Chin-Hui Lee[4],
Chenliang Xu[5], Rohit Kumar[5], Wei Chen[5], Jason Corso[5],
L. Fei-Fei[6], Daphne Koller[6], Vignesh Ramanathan[6], Kevin Tang[6], Armand Joulin[6], Alexandre Alahi[6]
[1] Kitware Inc, [2] Simon Fraser University, [3] Honeywell ACS Labs, [4] Georgia Institute of Technology, [5] University at Buffalo,
[6] Stanford University.

Abstract

Our MED 13 system is an extension of our MED 12 system [12, 13], and consists of a collection of low-level and high-level features, feature-specific classifiers built upon those features, and a fusion system that combines features both through mid-level kernel fusion and late fusion. Our MED submissions include total of 24 different configurations which consist of combinations of 2 submission timings (PS/AH), 3 training conditions (100/10/0Ex), and 4 types of feature conditions (Full/Visual/Audio/ASR).

Our MER 13 submissions reported recounting for all five MER events. Our MER system combines evidences from multiple base classifiers, which are translated to texts and used to identify key frames. Multiple MER results are fused and presented to users as recounting for each detection.

1 Introduction

For TRECVID 2013 [10, 15], we participated the Multimedia Event Detection (MED) and Multimedia Event Recounting (MER) tasks. For the MED task, we submitted runs on HAVIC dataset [16] for all the tasks, which include the pre-specified (E06–E15 and E21–E30), ad-hoc (E16–E20), and small example tests (10Ex).

Our MED 13 system is an extension of our MED 12 system [13, 8], and consists of a collection of low-level and high-level features, feature-specific classifiers built upon those features, and a fusion system that combines features both through mid-level kernel fusion and late fusion. Our MED submissions include total of 24 different configurations which consist of combinations of 2 submission timings (PS/AH), 3 training conditions (100/10/0Ex), and 4 types of feature conditions (Full/Visual/Audio/ASR).

Our MER 13 submissions reported recounting for all five MER events. Our MER system combines evidence from multiple base classifiers, which then are translated to text and used to identify key frames. Multiple MER results are fused and presented to users as recounting for each detection.

2 Multimedia Event Detection (MED)

Our MED results are summarized in Table 1. We observe that there is a substantial gap between 100Ex, 10Ex, and 0Ex systems. In terms of relative improvement from last year, our results from PS results indicate that our results on FullSys 100Ex runs are comparable to those on MED 2012 130Ex runs. Because 2012 results were based on 130Ex training condition, which has more of both positive and negative examples compared to this year's 100Ex, we can interpret the comparable numbers as an improvement in the 2013 system. In addition, we observed notable improvements in our 10Ex results from the last year.

While the MAP results presented in Table 1 are useful for understanding overall system performance, we also analyze the per-event average precision (AP) performance in Table 2 due to the significant event-to-event variation. Under to 0Ex condition, we see that many events perform quite poorly, but that a few events perform much better than the average. In particular, the 'changing a vehicle tire' and 'making a sandwich' events work quite well under 0Ex due to the performance of ASR features in this condition. This suggests that these events are particularly well-suited to a keyword spotting-type approach, and that they may have relatively higher audio quality than other events.

Comparing the 10Ex and 100Ex conditions, we see that the relative performance of different events is largely consistent. The best event under both conditions is 'flash mob gathering', and 'grooming an animal' is difficult under both conditions. However, it is surprising—at first—that the GENIE system performs better with 10 positive examples (10Ex condition) on 'making a sandwich' than it does with 100 positive

Table 1: MED results of mean average precision (%) on the PROGRESS Dataset (computed by NIST).

	MED'13			MED'12	
Condition	0Ex	10Ex	100Ex	10Ex	130Ex
Pre-specified FullSys	1.3	10.4	23.3	7.7	23.9
Pre-specified VisualSys	1.0	10.3	19.9	-	-
Pre-specified AudioSys	1.1	2.6	7.8	-	-
AdHoc FullSys	0.4	11.7	20.2	-	-
AdHoc VisualSys	1.2	7.3	16.9	-	-
AdHoc AudioSys	0.5	3.9	10.1	-	-

Table 2: Per-event results of average precision (%) on MEDTEST (self-reported).

	0Ex	10Ex	100Ex
Birthday party	7.1	13.8	34.3
Changing a vehicle tire	17.0	24.8	52.7
Flash mob gathering	1.5	56.4	74.3
Getting a vehicle unstuck	0.7	19.6	53.7
Grooming an animal	0.8	11.2	26.1
Making a sandwich	14.9	24.1	22.8
Parade	1.3	26.0	51.7
Parkour	1.0	41.0	57.4
Repairing an appliance	15.0	42.7	60.1
Working on a sewing project	4.5	10.5	28.1
Average (Mean AP)	6.4	27.0	46.1

examples. As explained in Section 2.1.4, though, this is due to differences in the fusion method, and the 0Ex ASR scores contribute relatively more under the 10Ex condition than under 100Ex.

We also note that our ‘expected performance’, in the sense of the mean AP on the MEDTEST data, is significantly better than the reported performance on the PROGRESS Dataset. We have no solid explanation for this difference. Two hypotheses are: (1) we performed unexpectedly poorly on the new events (E21–E30), causing a lower average; or (2) the positive-negative ratio in MEDTEST is significantly different in the PROGRESS Dataset. (Average precision is known to be sensitive to the positive-negative ratio in the test set, and will decrease when there are fewer positives even when the system does not change.)

2.1 10Ex and 100Ex

For 10Ex and 100Ex, we have improved our MED 12 system [8, 13] by adding a set of new features and enhancing fusion methods. We basically learned base classifiers separately by using a single feature or a subset of features, then fused them with a late fusion method.

2.1.1 Features

In our MED 13 system, we computed a set of features, which are mostly quantized by a codebook-based method. For many features, a single clip-level histogram representation based on bag-of-words (BoW) models were used, while a sequence of BoW segments were built for ObjectBank features to be incorporated into temporal latent SVM models.

The list of our visual features includes HOG2x2, HoG3D [4] with varying spatio-temporal quantization, a set of ObjectBank [7] variants, Scene Attributes [11], GIST [9], Color SIFT [18], independent subspace analysis (ISA) [5], transformed color histogram [18], dense SIFT, Sparse SIFT with Hessian and MSER interest point detection, dense trajectory, SSIM, Self similarity, Geometric context color histogram, and Action Bank [14].

Additionally, the list of our audio features include MFCC [6], acoustic segment models (ASMs) [1], Audio Bank [3], and words from automatic speech recognition (ASR) computed and shared by BBN and SRI Menlo Park.

2.1.2 Training Protocol

For event agent generation, following TRECVID guidelines, we have followed an independent event agent generator scenario where only the corresponding event kit examples are used along with Background (BG) dataset. In other words, the other remaining event kits were assumed to be unknown and not included in any way (e.g., as negative examples).

For the incorporation of negative samples beyond event kit samples, we have taken a realistic set-up where we assume that only the labels from the event kits are known and treat all the remaining BG datasets to be unknown. In our approach, we use the exemplars in the BG datasets as negative training data (again, regardless of their labels), which simulates a realistic scenario where the negative datasets are essentially polluted.

Specifically, for the agent generator for a given event class, our training data consists of the samples in the corresponding event kit and samples in the BG datasets. For the examples included in the event kits as 'near miss' or 'related', they were used as explicitly as negative samples while the remaining 'positive' samples from the event kit are used as positive training data. Any samples drawn from the BG datasets were treated as negative samples, regardless of their actual content.

2.1.3 Base Classifiers

From the feature sets, we learned multiple classifiers from different subsets of features, which compute scores on the test dataset independently. These set of classifiers are called 'base classifiers' in our framework. Most base classifiers are learned from single features, while some are learned from multiple features jointly. In total, we learned 35 base classifiers.

Two different types of base classifiers were learned: (1) non-linear kernel SVMs using kernels [2] such as histogram intersection kernel and (2) multiple kernel learning (MKL) across multiple features by using the AND/OR graph structure [17]. During the testing on the PROGRESS dataset, the base classifiers are applied to the test data and independently compute detection scores for the corresponding event class. Accordingly, each test clip is associated with multiple base classifier scores, which are then fused to compute final scores.

2.1.4 Fusion

Throughout the GENIE system, we use geometric mean to fuse the scores of various base level classifiers, as the geometric mean has two key advantages. First, it does not require training data, which is important because—in the 10Ex case—all positive examples are used for base classifier training. Second, fusing base classifier scores with the geometric mean has been found empirically to provide roughly monotonically-increasing MAP as base classifier scores are added, provided that the base classifier scores have similar distributions.

Figure 1 illustrates this point with the mean average precision (MAP) of incrementally fusing base classifiers in decreasing order of performance. Both the base classifier performance and the fused MAP are computed on events 6–15 in the MEDTEST data set; we exclude events 21–30 as we suspect that there are unlabeled positive instances of these events in MEDTEST and, anyway, because the small number of labeled positives adds some noise to our measured performance. This progression also shows that there are two notable regions of rapid improvement. The first takes place during the fusion of the first few base classifiers; these highest-performing base classifiers are various visual features. The second, more gradual improvement in fused MAP occurs starting with the 25th base classifier, and illustrates the improvement gained by adding audio information. While the first 25 base classifiers contain visual data—each of which is higher-performing

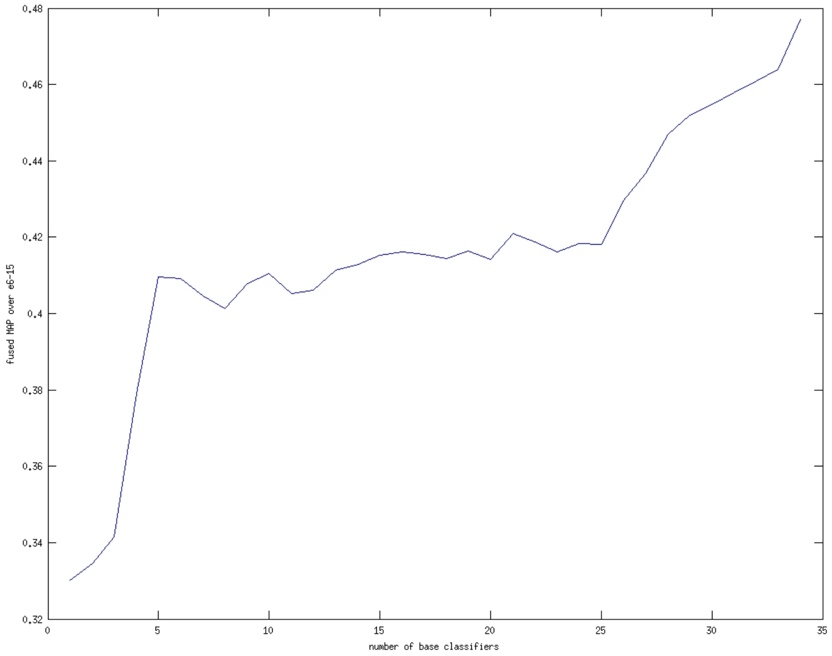


Figure 1: Mean Average Precision (MAP) over events 6–15 (measured against MEDTEST) as a function of the number of base classifiers fused using geometric mean. While there are relatively small decreases along the way, MAP is roughly monotonically increasing with the number of base classifiers.

than the best audio classifier—the relative contribution to fusion of the audio features is higher because they bring complementary information.

Whereas the 100Ex fusion is a straightforward geometric mean over all of the base classifiers—with some complexity needed to handle those clips with missing base classifier scores—the 10Ex fusion method is somewhat more elaborate. Instead of a single-level fusion, we perform a multi-level fusion. Base classifiers are divided into three categories:

1. Audio features, such as low-level MFCC features, ASM, and mid-level Audio Bank scores;
2. Color features, such as TCH and CSIFT; and
3. Low-level motion and object features, such as ObjectBank and various gradient histogram features.

At the first level, all of the base classifiers in each category are fused using the geometric mean. Then, these three scores are further fused via geometric mean for all event categories. Finally, for the features where we have found high ASR performance in the 0Ex condition—making a sandwich and changing a vehicle tire, specifically—we finally fuse a third time. We have found that fusion of the 0Ex ASR scores improve the AP of these events by 5–10%, but that they reduce AP on other events.

2.2 0Ex

The MED13 0Ex task involves searching a large video archive for specific event in the absence of any training videos. The search needs to be based simply on the textual event description. This necessitates the association of video clips with human-nameable semantic concepts.

To drive our 0Ex system, we use semantic video and audio features. For visual features we employ ObjectBank and Scene Attributes which were described in the Section 2.1.1. Both of these features are

Table 3: MER results for the GENIE system

Accuracy	ObsTextScore	PRRT
56.44%	0.9	141.23%

average pooled across frames from a given clip to obtain a clip-level score. With 177 dimensional ObjectBank features and 102 dimensional scene attributes, we obtain a bank of about 279 semantic visual concepts which we can use for the zero-shot search. For zero shot search to work effectively, it is important for the concept scores to be meaningful to a human and have a direct correlation to the semantics that are observed in the video.. For example, a score of 0.5 from a person detector and 0.5 from a water detector may not imply an equal probability of these concepts being present in the clip. In order to convert the raw detector scores into meaningful values, we employ a score calibration scheme to modify the concept scores.

The semantic concepts which are relevant to the keywords in an event kit description are used to construct a zero shot query. The search query is formulated in the form of an AND-OR query. An AND operation tries to determine the likelihood of multiple concepts being present simultaneously in a video clip. AND operation over two or more concepts averages the respective scores. An OR operation tries to detect the presence of at least one of the requested concepts by finding the maximum of the corresponding scores. The scores returned after evaluating the complete query for each event give the scores for each of the test clips. For capturing the semantics in the audio stream, we used Automatic Speech Recognition (ASR). The scores obtained from these three different features (object bank, scene attributes, and ASR) are combined to obtain final test scores, in a manner similar to 10Ex and 100Ex systems.

3 Multimedia Event Recounting (MER)

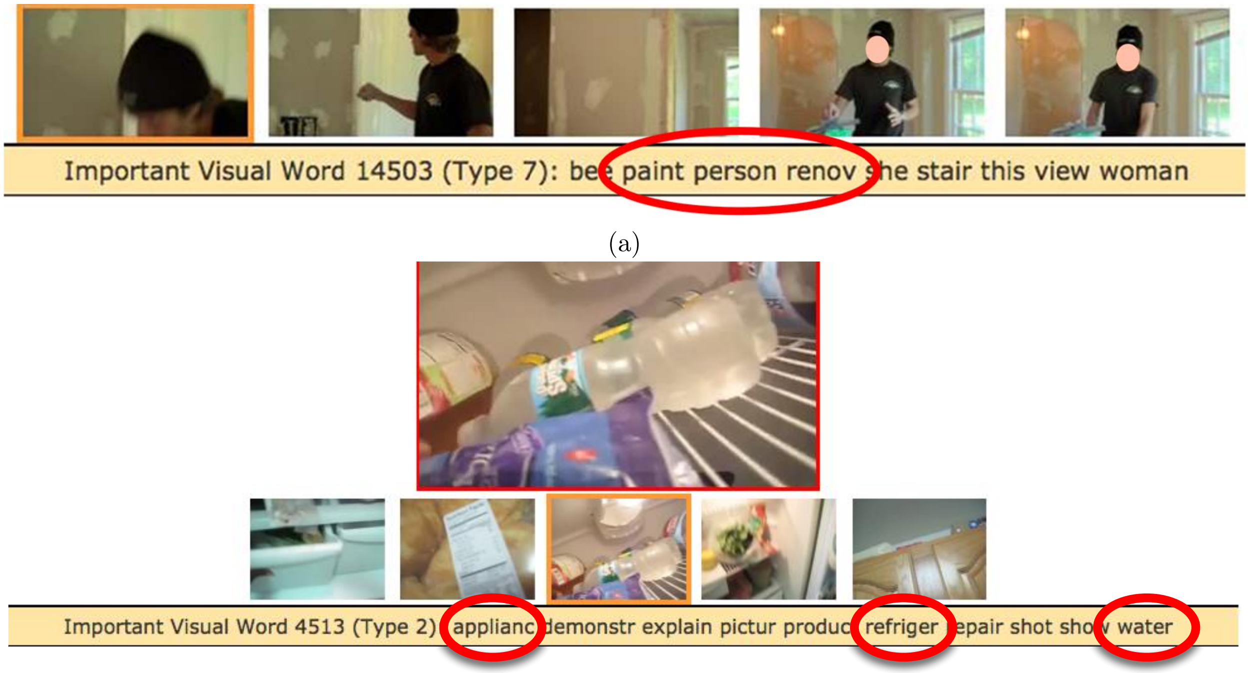
Our MER results are summarized in Table 3. Overall, the results indicate that our MER system shows reasonable accuracy. However, based on the ObsTextScore and PRRT metrics, it outputs a lot of information; this will be reduced in future versions of the system.

In order to recount the classifiers based on low-level features, we used a topic model to 'translate' the feature into text. In particular, we extracted the top 1000 dimensions for each feature (using dimensionality reduction), and assigned a translation based on a previously learned topic model. The topic models were learned on non-HAVIC data downloaded from the Internet. For the semantic features (like Object Bank), we simply used the detector labels as the translation. In addition to the text translation, we also located the most meaningful frames for each classifier.

Figure 2(a) shows the visualization of sparse SIFT feature on clip HVC018429 with event E026 ('renovating a home'). The words circled in red show words that are particularly suited to summarizing the extracted frames (and the video clip). Figure 2(b) shows the visualization of HOG3D feature on clip HVC116421 with event E022 ('cleaning an appliance'). We can see the topic model translation indeed found a few meaningful words 'applianc', 'refriger' and 'water', which are circled in red in the figure. Also, the extracted frames are very good match to the text descriptions.

4 Conclusion

In our MED 13 system, we have explored the use of a large number of features and advanced event agent learning methods which expanded our system beyond our MED 12 system. Meaningful improvements were achieved in our 10Ex performance. We also observed comparable 100Ex results when compared to our MED 12 results, which can be interpreted as an improvement considering the number of positive and negative samples notably reduced from the last year. The improvement in our performance indicates that the overall direction of our research and development addresses the problem more effectively. We obtained these results in spite of the minimal optimization and tuning efforts, and clearly indicating the reliability of the developed


Important Visual Word 14503 (Type 7): be paint person renov he stair this view woman
Important Visual Word 4513 (Type 2): applianc demonstr explain pictur produc refriger repair shot show water

(a)

(b)

Figure 2: Example of MER visualization for (a) sparse SIFT on HVC018429 with E026 (‘renovating a home’) and (b) HOG3D on HVC116421 with E022 (‘cleaning an appliance’).

system. In the future, we plan to tie our MED and MER system to simultaneously improve the performance for both tasks while enhancing the transparency of the decision criteria made by our system.

5 Acknowledgment

We thank BBN and SRI Menlo Park for sharing their ASR features computed on HAVIC dataset.

This work is supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center contract number D11PC20069. 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.

References

[1] B. Byun, I. Kim, S. M. Siniscalchi, and C.-H. Lee. Consumer-level multimedia event detection through unsupervised audio signal modeling. In *Interspeech*, 2012.

[2] C.-C. Chang and C.-J. Lin. Libsvm: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.*, 2:27:1–27:27, May 2011.

[3] Z. Huang, Y.-C. Cheng, K. Li, V. Hautamaki, and C.-H. Lee. A Blind Segmentation Approach to Acoustic Event Detection Based on I-Vector. In *INTERSPEECH*, 2013.

[4] A. Kläser, M. Marszalek, and C. Schmid. A spatio-temporal descriptor based on 3d-gradients. In *BMVC*, 2008.

[5] Q. Le, W. Zou, S. Yeung, and A. Ng. Learning hierarchical spatio-temporal features for action recognition with independent subspace analysis. In *CVPR*, 2011.

[6] C.-H. Lee, F. Soong, and B.-H. Juang. A segment model based approach to speech recognition. In *ICASSP*, 1988.

[7] L.-J. Li, H. Su, E. P. Xing, and L. Fei-Fei. Object Bank: A High-Level Image Representation for Scene Classification and Semantic Feature Sparsification. In *Proceedings of the Neural Information Processing Systems (NIPS)*, 2010.

[8] S. Oh, S. McCloskey, I. Kim, A. Vahdat, K. Cannons, H. Hajimirsadeghi, G. Mori, A. Perera, M. Pandey, and J. J. Corso. Multimedia event detection with multimodal feature fusion and temporal concept localization. *Machine Vision and Applications*, July 2013.

[9] A. Oliva and A. Torralba. Modeling the shape of the scene: a holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42(3):145–175, 2001.

[10] P. Over, G. Awad, M. Michel, J. Fiscus, G. Sanders, W. Kraaij, A. F. Smeaton, and G. Queenot. Trecvid 2013 – an overview of the goals, tasks, data, evaluation mechanisms and metrics. In *Proceedings of TRECVID 2013*. NIST, USA, 2013.

[11] G. Patterson and J. Hays. Sun attribute database: Discovering, annotating, and recognizing scene attributes. In *Proceeding of the 25th Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[12] A. Perera, S. Oh, M. Leotta, I. Kim, B. Byun, C.-H. Lee, S. McCloskey, J. Liu, B. Miller, Z. Huang, A. Vahdat, W. Yang, G. Mori, K. Tang, D. Koller, L. Fei-Fei, K. Li, G. Chen, J. Corso, Y. Fu, and R. Srihari. GENIE TRECVID 2011 Multimedia Event Detection: Late-Fusion Approaches to Combine Multiple Audio-Visual Features, 2011.

[13] A. Perera, S. Oh, M. Pandey, T. Ma, A. Hoogs, A. Vahdat, K. Cannons, H. Hajimirsadeghi, G. Mori, S. McCloskey, B. Miller, S. Venkatesha, P. Davalos, P. Das, C. Xu, J. Corso, R. Srihari, I. Kim, Y.-C. Cheng, Z. Huang, C.-H. Lee, K. Tang, L. Fei-Fei, and D. Koller. TRECVID 2012 GENIE: Multimedia Event Detection and Recounting, 2012.

[14] S. Sadanand and J. J. Corso. Action bank: A high-level representation of activity in video. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 1234–1241. IEEE, 2012.

[15] A. F. Smeaton, P. Over, and W. Kraaij. Evaluation campaigns and trecvid. In *MIR '06: Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval*, pages 321–330, New York, NY, USA, 2006. ACM Press.

[16] S. Strassel, A. Morris, J. Fiscus, C. Caruso, H. Lee, P. Over, J. Fiumara, B. Shaw, B. Antonishek, and M. Michel. Creating havic: Heterogeneous audio visual internet collection. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, may 2012.

[17] K. Tang, B. Yao, L. Fei-Fei, and D. Koller. Combining the right features for complex event recognition. In *CVPR*, 2013.

[18] 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*, 32(9):1582–1596, 2010.