TRECVID 2012 GENIE:
Multimedia Event Detection and Recounting

A. G. Amitha Perera[1] , Sangmin Oh[1], Megha Pandey[1], Tianyang Ma[1], Anthony Hoogs[1], Arash Vahdat[2], Kevin Cannons[2], Hossein Hajimirsadeghi[2], Greg Mori[2], Scott McCloskey[3], Ben Miller[3], Sharath Venkatesha[3], Pedro Davalos[3], Pradipto Das[4], Chenliang Xu[4], Jason Corso[4], Rohini Srihari[4], Ilseo Kim[5], You-Chi Cheng[5], Zhen Huang[5], Chin-Hui Lee[5], Kevin Tang[6], L. Fei-Fei[6], Daphne Koller[6]

Abstract

Our MED 12 system is an extension of our MED 11 system [12], 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 score fusion. We have incorporated large number of audio-visual features in our new system and incorporated diverse types of standard and newly developed event agents which learn the salient audio-visual characteristics of event classes. The combination of additional features and newly developed powerful event agents improve our MED performance substantially beyond our MED 11 results.

In addition, our MER 12 submissions reported recounting of specified clips for all five MER events and additionally provided MER results for all the clips detected by MED system. Our MER system generated recounting of detections based on CDR features and synopsis provided as part of the EventKits and DEV-T datasets. The MER evaluation results are promising for event-level discrimination, and indicated further improvement to be made for clip-level discrimination.

1 Introduction

For TRECVID 2012 [11], we participated MED and MER tasks. For MED task, we have submitted runs for all enlisted tasks which include pre-specified (E06-E15 & E21-E30), ad-hoc (E16-E20), and small example tests (Ex10).

Our MED 12 system is an extension of our MED 11 system [12], 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 score fusion. We have incorporated additional features in addition to the ones used for MED 11. In addition, we have incorporated additional event agents which learn the salient visual characteristics through latent SVM framework. The combination of additional features and newly developed powerful event agents improve our MED performance substantially beyond our last year's results.

Our MER 12 submissions reported recounting of specified clips for all five MER events and additionally provided MER results for all the clips detected by MED system. Our MER system generated recounting of detections based on CDR features and synopsis provided as part of the EventKits and DEV-T datasets. The MER evaluation results were very promising for event-level discrimination, and indicated further improvement to be made for clip-level discrimination.

2 Multimedia Event Detection

For TRECVID MED 12, we have improved our MED 11 system [12] by adding a set of new features and enhancing fusion methods.

Overall, our goal for MED task was to optimize the event agents to produce performance pertaining to the ratio of 1:12.5 between probability of detection and false alarms. The averaged results reported by NIST is summarized in Table 1 where it can be observed that the actual average ratio is fairly close to the ratio of 12.5, which is encouraging.

1 Kitware Inc. 2 Simon Fraser University. 3 Honeywell ACS Labs. 4 University at Buffalo. 5 Georgia Institute of Technology. 6 Stanford University.

	Probability of detection (Pd)	False Alarm Ratio (FAR)	Ratio of Pd/FAR
Pre-specified Full	0.0261	0.3346	12.82
Ad-Hoc Full	0.0294	0.3134	10.66
Pre-specified EK10Ex	0.0392	0.5217	13.30
Ad-Hoc EK10Ex	0.0393	0.5790	14.73

Table 1: Results of average performance on Progress Dataset (computed by NIST)

In addition, we pursued an event-agnostic approach for MED, which means that we made no effort to tune our system for particular events, even for pre-specified events, except in an unsupervised manner during the event generation steps. It is also worth noting that we conducted only minimal tuning effort for CDR generation such as codebook generation or event agent parameter estimation. Still, the effect of adding features and employing advanced fusion methods resulted in fairly promising performance, which is encouraging.

In particular, the effect of the improvement made for our system can be observed by comparing the estimated DET curves generated during cross validation, which are shown side-by-side in Figure 1. Additionally, the estimated DET curves for MED Full tasks for E16-E20 and E21-E30 are shown in Figure 2, which show fairly comparable and reliable performance for E06-E15 shown in Figure 1(Right). The overall difference between Pre-speicified and Ad-hoc events are not noticeable, which indicates the reliability of the developed system.

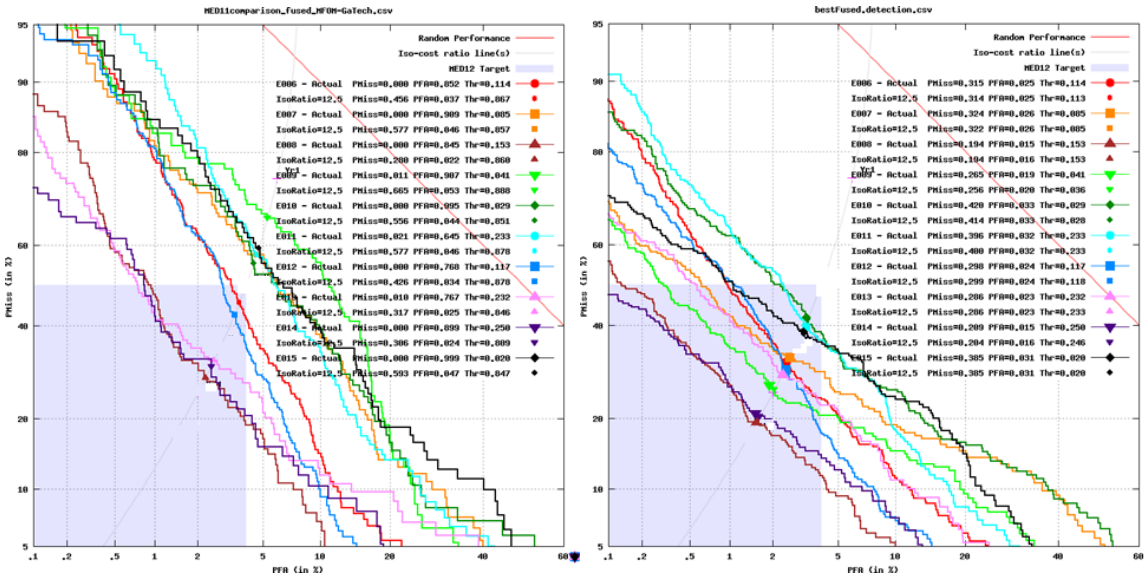

Figure 1: DET curves for E06-E15 estimated via cross validation on training dataset using identical data splits: (Left) DET curves by MED 11 system. (Right) DET curves by MED 12 system. The improvement in performance towards MED 12 system is clearly visible. (Note that the MED 11 curves on the left are different from the curves in[12] because the training protocol for MED 12 was different than MED 11. The curve here reflects the MED 11 system re-trained with the new protocol.)

2.1 Features

In our MED 12 system, we computed a set of features to construct our CDR, 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.

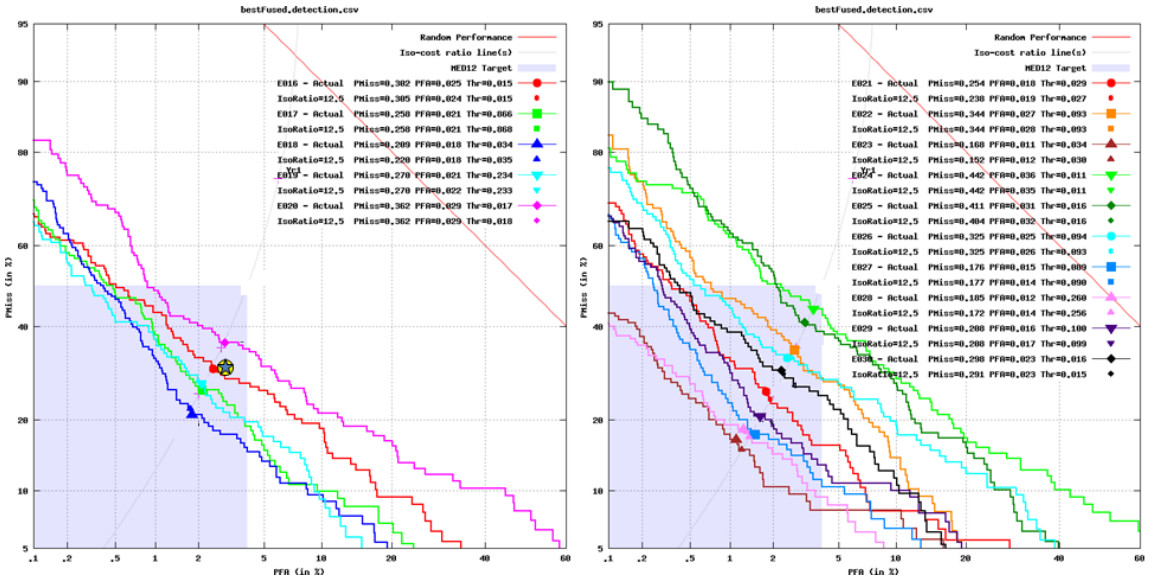

Figure 2: DET curves for Full MED setting, estimated via cross validation on training dataset using identical data splits: (Left) DET curves for E16-E20, and (Right) DET curves for E21-E30.

The list of our visual features includes HoG3D [5], ObjectBank [8], GIST [10], Color SIFT [13], independent subspace analysis (ISA) [6], transformed color histogram [13], and a set of visual features from [15] including histogram of gradients, geometry texton histogram, self-similarity measure, dense/sparse SIFT, local binary patterns (LBP), tiny image, and etc.

Additionally, the list of our audio features include MFCC [7] and acoustic segment models (ASMs) [1].

2.2 Protocol for Training

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 DEV dataset. In other words, the other remaining event kits were assumed to be unknown and not included in any way (e.g., as negative datasets).

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 DEV datasets to be unknown, i.e., as negative samples. Accordingly, DEV datasets are always used as negative training data regardless of their labels, which simulate a realistic scenario where the negative datasets are essentially polluted.

Specifically, for each event agent generator for every event class, our training data consists of the samples in the corresponding event kit and samples in the DEV datasets (including MED11TEST) from MED '11. 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 DEV datasets were treated as negative samples, even if they were labeled as positive for the target event class; i.e., we ignored the labels in the DEV datasets, which is a realistic scenario.

2.3 Base Classifiers

From the CDR feature sets, multiple classifiers are learned from different subsets of features, which compute scores on PROGRESS dataset independently. These set of classifiers are categorized as 'base classifiers' in our framework. A large number of base classifiers are learned from single features, while others are learned from multiple features jointly. In total, we learned 14 base classifiers.

We learned three different types of base classifiers: (1) non-linear kernel SVMs using kernels [2] such as histogram intersection kernel, (2) multiple kernel learning (MKL) [14] for joint event agent learning across

multiple features, and (3) a temporal variation of latent SVM.

In particular, our temporal variant of latent SVM is a new development which learns salient parts of videos and detect such important temporal regions from test clips, resulting in significant boost in performance. The overall formulation is analogous to the original formulation of spatial latent SVM model for object detection [3]. The training and testing of temporal latent SVM is computationally more demanding than standard non-linear kernel SVMs, and it has been employed for ObjectBank feature only for MED 12.

During the testing on 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 fused to compute final scores.

2.4 Fusion

To compute the final single number score for detection, multiple base classifier scores are fused through diverse fusion methods. We have learned fusion parameters through cross-validation on our training data when learning is needed, or used blind fixed-rule fusion methods. By blind fusion methods, we mean simple fixed approaches such as average and geometric mean of base classifier scores. In terms of learning-based parameter estimation for fusion methods, we have used multiple methods including Expert Forest [9], MFoM [4], linear SVM, etc.

The learning of fusion classifiers is conducted via cross-validation on training data where it has been observed that the best performing fusion method can be different across events, with significant differences in the performance of across the set of methods. Overall, geometric mean and expert forest are observed to be selected frequently as the best performing fusion methods for largest number of events. MFoM is also observed to be good for a fair number of events, while linear SVM and average were selected only for a small number of classes.

Once multiple fusion models are established, the best fusion classifier is selected per event class based on performance estimated by cross validation on training data. The selected method was then used to fuse scores on the test dataset to generate final scores.

3 Multimedia Event Recounting

The task of MER is to report the list of observations and rationale for each and every MED detection. Our MER submissions include: (1) detailed MER versions for pre-specified list of MER test video clips, and (2) compact version of MER results for all the video clips detected by our MED system. A sample snapshot of an automatically generated output by our MER system is shown in Figure 3.

In particular, our MER system encompasses models for human language and the meaning of text words (both nouns and verbs) included in Event Kits. For example, it can be observed in Figure 3 that our MER outputs include the key nouns and verbs along with their definitions.

Specifically, given a detected clip, our MER system identifies the best matching clip from EventKit samples based on multiple low-level CDR features and semantic object detections. Then, the LDC-provided synopses for the matched clips in the training data are used in a data-driven manner to describe the target video clip, where the confidence scores are generated based on the likelihood of the matches.

Based on evaluation by NIST, our MER system showed the best results for the task of event discrimination among all submissions, but, indicated a room for improvement for clip-level differentiation. The outcome can be attributed to the synopsis transfer methodology adopted in our system. We plan to improve the clip-level differentiation capability by extending multiple aspects of our current system, including the semantic detection of salient per-clip characteristics and language models.

4 Conclusion

In our MED 12 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 11 system. The improvement in our performance is significantly noticeable and 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

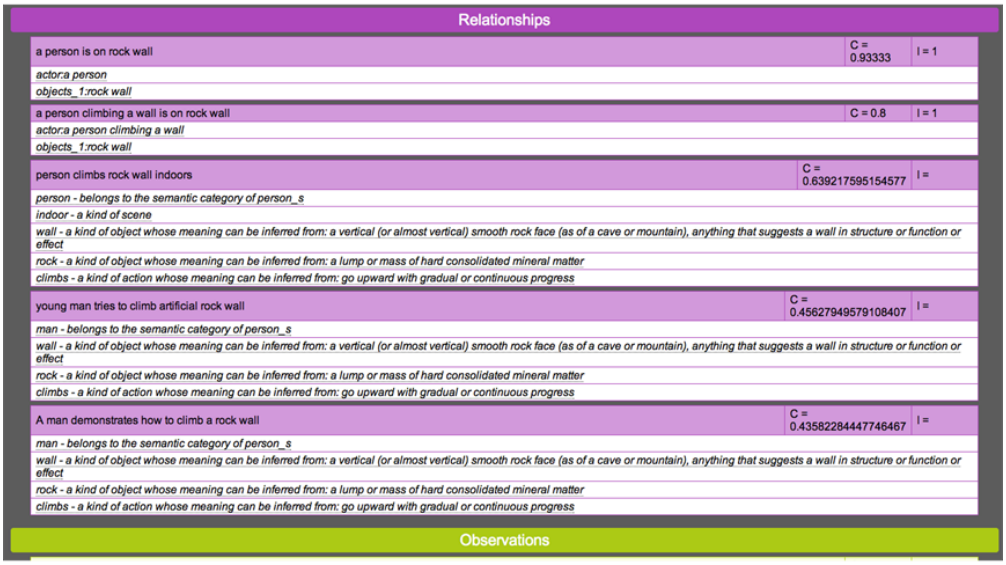

Figure 3: An example snapshot of an output by our MER system.

tuning efforts, and clearly indicating the reliability of the developed 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

This work was 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 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] Byungki Byun, Ilseo Kim, S. M. Siniscalchi, and Chin-Hui Lee. Consumer-level multimedia event detection through unsupervised audio signal modeling. In *Interspeech*, 2012.

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

[3] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1627–1645, 2010.

[4] Ilseo Kim, Sangmin Oh, Byungki Byun, A.G. Amitha Perera, and Chin-Hui Lee. Explicit performance metric optimization for fusion-based video retrieval. In *Workshop on Information Fusion in Computer Vision for Concept Detection, in conjunction with European Conference on Computer Vision (ECCV)*, 2012.

[5] Alexander Kläser, Marcin Marszalek, and Cordelia Schmid. A spatio-temporal descriptor based on 3d-gradients. In *BMVC*, 2008.

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

[7] Chin-Hui Lee, F.K. Soong, and Bing-Hwang Juang. A segment model based approach to speech recognition. In *ICASSP*, 1988.

[8] Li-Jia Li, Hao Su, Eric P. Xing, and Li 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.

[9] Jingchen Liu and Scott McCloskey. Local Expert Forest of Score Fusion for Video Event Classification. In *ECCV*, 2012.

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

[11] Paul Over, George Awad, Martial Michel, Jonathan Fiscus, Greg Sanders, Barbara Shaw, Wessel Kraaij, Alan F. Smeaton, and Georges Quéenot. Trecvid 2012 – an overview of the goals, tasks, data, evaluation mechanisms and metrics. In *Proceedings of TRECVID 2012*. NIST, USA, 2012.

[12] A.G.A. Perera, S. Oh, M. Leotta, I. Kim, B. Byun, C-H. Lee, S. McCloskey, J. Liu, B. Miller, Z.F. 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] Koen E. A. van de Sande, Theo Gevers, and Cees G. M. Snoek. Evaluating Color Descriptors for Object and Scene Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1582–1596, 2010.

[14] A. Vedaldi, V. Gulshan, M. Varma, and A. Zisserman. Multiple Kernels for Object Detection. 2009.

[15] J. Xiao, J. Hays, K. Ehinger, A. Oliva, and A. Torralba. SUN Database: Large-scale Scene Recognition from Abbey to Zoo. 2010.