GENIE TRECVID2011 Multimedia Event Detection: Late-Fusion Approaches to Combine Multiple Audio-Visual features

A. G. Amitha Perera[1], Sangmin Oh[1], Matthew Leotta[1], Ilseo Kim[2], Byuungki Byun[2], Chin-Hui Lee[2],Scott McCloskey[3], Jingchen Liu[3], Ben Miller[3], Zhi Feng Huang[4], Arash Vahdat[4], Weilong Yang[4], Greg Mori[4], Kevin Tang[5], Daphne Koller[5], L. Fei-Fei[5], Kang Li[6], Gang Chen[6], Jason Corso[6], Yun Fu[6], Rohini Srihari[6]

Abstract

For TRECVID 2011 MED task, the GENIE system incorporated two late-fusion approaches where multiple discriminative base-classifiers are built per feature, then, combined later through discriminative fusion techniques. All of our fusion and base classifiers are formulated as one-vs-all detectors per event class along with threshold estimation capabilities during cross-validation. Total of five different types of features were extracted from data, which include both audio or visual features: HOG3D, Object Bank, Gist, MFCC, and acoustic segment models (ASMs). Features such as HOG3D and MFCC are low-level features while Object Bank and ASMs are more semantic. In our work, event-specific feature adaptations or manual annotations were deliberately avoided, to establish a strong baseline results. Overall, the results were competitive in the MED11 evaluation, and shows that standard machine learning techniques can yield fairly good results even on a challenging dataset.

Summary of Submitted Runs

1. `GENIE_MED11_MED11TEST_MEDFull_AutoEAG_p-MFoMaudiovisual_1`: Fusion by MFoM with 5 audio-visual features
2. `GENIE_MED11_MED11TEST_MEDFull_AutoEAG_c-MLE_1`: Fusion by MLEs with 5 audio-visual features
3. `GENIE_MED11_MED11TEST_MEDFull_AutoEAG_c-MFoMvideo_1`: Identical to #1, with only 3 video features
4. `GENIE_MED11_MED11TEST_MEDFull_AutoEAG_c-MFoMaudio_1`: Identical to #1, with only 2 audio features

1 Introduction

For TRECVID 2011 MED task, GENIE system incorporated two late-fusion approaches where multiple discriminative base-classifiers are built per feature, then, combined later through discriminative fusion classifiers. All of our fusion and base classifiers are formulated as one-vs-all detectors per event class along with threshold estimation capabilities during cross-validation.

Our goal was to establish a baseline performance result using state-of-the-art techniques that already exist in the computer vision and multimedia retrieval communities. To that end, we focused on fully automatic computing, and deliberately avoided tuning to the MED11 data, such as adapting features to the MED11 data or training new features based on manual annotations. We did, of course, train classifiers over these features using the MED11 training data. Overall, the results were competitive in the MED11 evaluation, and shows that standard machine learning techniques can yield fairly good results even on a challenging dataset.

Our results were based on five audio and visual feature types (HOG3D, Object Bank, Gist, MFCC, and ASMs), as described in Section 2. We trained discriminative, one-vs-all classifiers based on each feature type for each event. For the 15 MED events, we therefore had 45 base classifiers. Then, these base classifier scores were fused to generate the final 15 event classifiers. We explored two different fusion schemes—a single global classifier and a mixture of local experts—as detailed in Section 3.

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

2 Features and Base Classifiers

We extracted five types of features from data. They include both audio and visual features at different granularities: HOG3D bag-of-words (BoW), Object Bank, Gist, MFCC BoW, and acoustic segment models (ASMs) BoW. These features are shown in Table 1.

Overall, it took approximately 5700 hours in total to extract all five features, where the time taken for each feature type is shown in Table 1. The majority of feature extraction time was spent for computationally intensive visual features, while audio features are computed in a relatively short period of time. The most computationally intensive feature is Object Bank. To reduce computation time, we only applied Object Bank on key frames, where the key frames were determined by looking at changes in overall RGB color histograms of the frame. The key frame algorithm took approximately 100 hours for entire MED11 dataset.

Once clip-level features were computed, standard SVM classifiers were learned in one-vs-all manner for all event classes. In each case, all the exemplars from the corresponding event kit were used as positive samples, and the exemplars from the other event kits were used as negative samples. These per-feature classifiers constitute the base classifiers of our system. Later, as described in Section 3, base classifier scores on test data are fed into fusion classifiers to generate final scores.

Details of each feature and corresponding base classifier performance are described in the following sections.

	Category	Granularity	Time (hours)	Temporal sampling
HOG3D BoW	Visual	low-level	1426	uniform
Object Bank	Visual	high-level	3250	key frames
Gist	Visual	high-level	700	uniform
MFCC BoW	Audio	low-level	30	none
ASM BoW	Audio	high-level	60	none

Table 1: Five different types of audio-visual features and their granularities and processing time (hours): HOG3D bag-of-words (BoW), Object Bank, Gist, MFCC BoW, and ASM BoW.

2.1 HOG3D

HOG3D [5] is a variant of popular histogram of gradients (HoG) descriptor, with additional information using the temporal aspect of videos. In our work, we resized the video frames such that the largest dimension (height or width) was 160 pixels, and extracted densely sampled HOG3D features. Our HOG3D parameters resulted in a 300-dimension feature vector. We then used K-means clustering to create a 1000-word codebook from a random sampling of the event kit data. Finally, we pooled all the quantized HOG3D features across the video to create a single HOG3D bag-of-words (BoW) model.

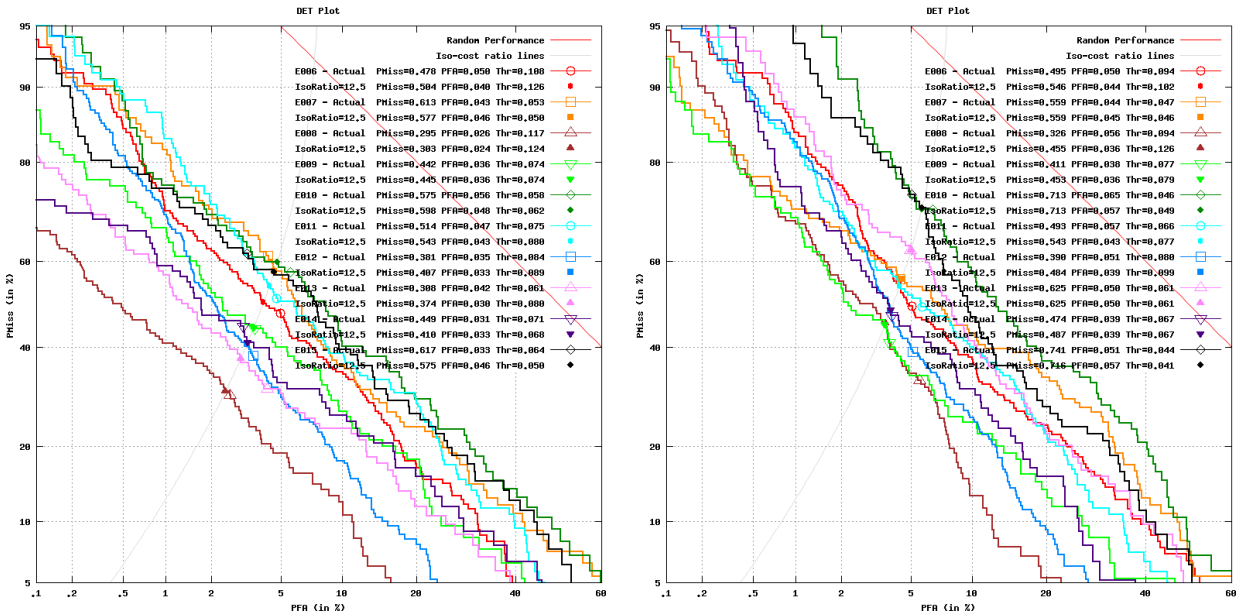
On top of this HOG3D BoW feature, we trained a non-linear SVM [1] using a histogram intersection kernel (HIK) distance function. Our experiments with several kernel types showed better performance for HIK kernels for all bag-of-words features. In particular, HIK kernels demonstrated clear advantage over linear kernels for SVMs.

The performance of HOG3D features for all MED11 event classes on official test dataset are shown in Figure 1(a).

2.2 Object Bank

Object Bank [6] consists of a set of detectors for a large array of distinct objects. Our existing Object Bank implementation [6] already incorporates detection modules for 171 object classes and is one of the few (if not only) integrated large-scale object recognition system publicly available. The Object Bank system is arbitrarily expandable and open, which means that, regardless of object classes, every object detector is trained in a generic framework and can be easily plugged-in into the main system.

The overall computational process of Object Bank is shown in Figure 2: an array of object detectors are applied to images at various scales, and their responses are recorded along with spatial layout information



(a) HOG3D (b) Object Bank

Figure 1: DET curves for base classifiers per feature type on MED11 Test dataset, for all ten event classes.

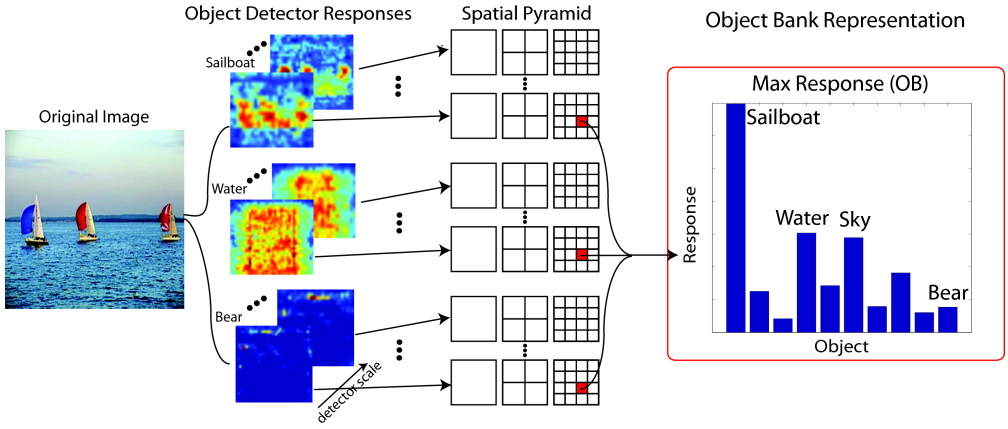


Figure 2: Scene descriptor based on Object Bank responses. Multiple semantic object detector responses are recorded and concatenated w.r.t. spatial layout to form a scene appearance descriptor. (Reproduced from [6].)

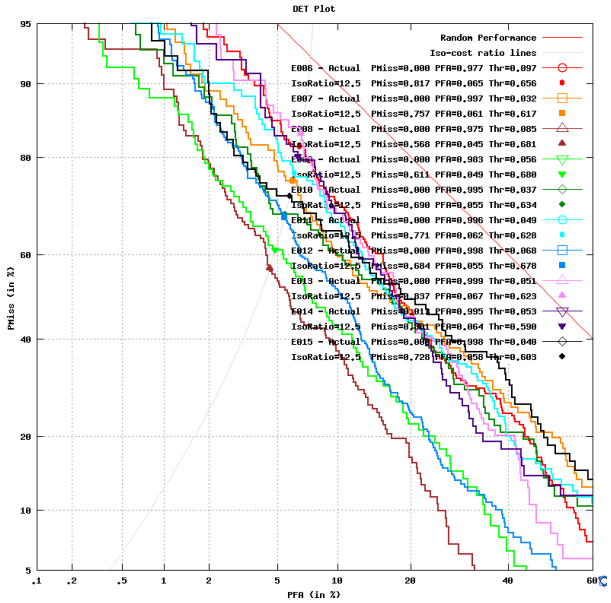


Figure 3: DET curves for Gist base classifier on MED11TEST data.

to form high-dimensional scene appearance descriptors. Object Bank is applied on a per image basis, and multiple Object Bank features are pooled to produce a clip-level descriptor across frames. In our experiments, we simply took the max responses from each Object Bank feature dimension across key frames.

It is worth noting that the responses of object detectors employed within Object Bank are still noisy. In particular, many of the MED data clips have low-resolution quality or present severe degree of motion blur. All these factors compound together and generate noisy Object Bank feature vector.

The performance of an SVM classifier on Object Bank features for all MED11 event classes are shown in Figure 1(b).

2.3 Gist

In order to exploit correlations between events and the environments in which they take place, our system incorporates the Gist feature [8]. Gist features represent an image's content in particular spatial frequency bands, and has been shown to provide discrimination between different types of environments such as natural versus man-made, open (i.e. outdoor) versus closed (indoor). It is applied as a full-frame descriptor. That is, we compute a single Gist feature for a given video frame. We extracted Gist features for every 10th frame of video clips.

Given the per-frame nature of the Gist feature, there are several alternatives to generating a clip-level classification. We chose a per-frame classification approach, where each frame is classified independently, and these are combined to form a per-clip probability.

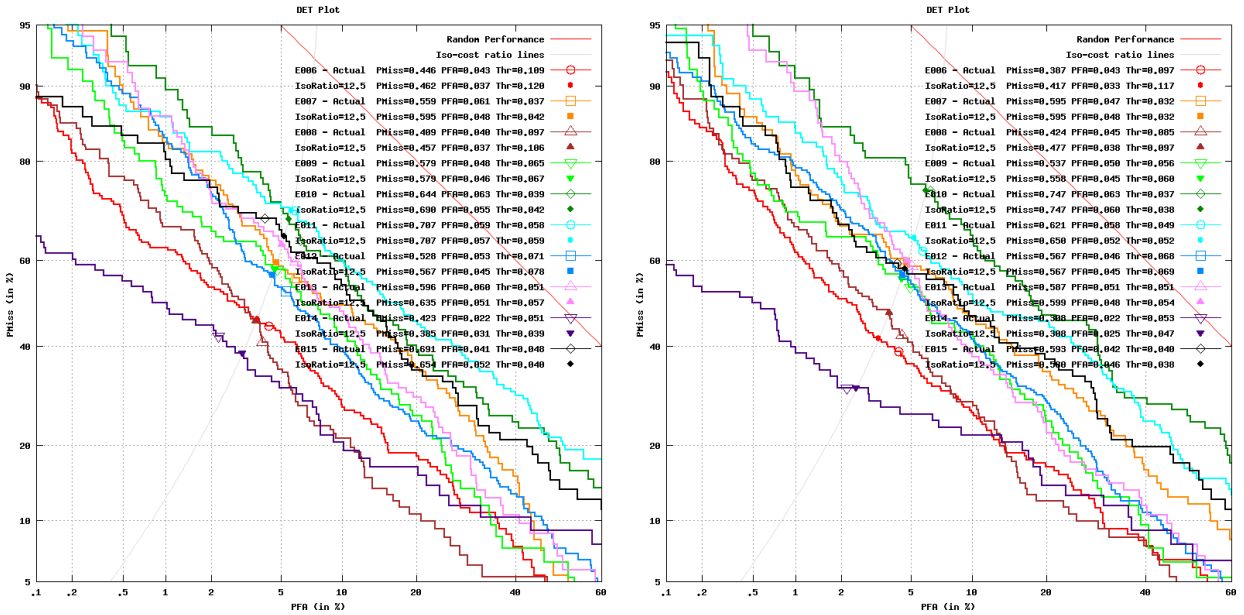
In more detail: we train a bank of one-versus-all linear SVM classifiers, using libSVM [1], for each of the 10 evaluation events. The training set consists of data from the event kits, as well as DEV-T. On unlabeled clips, Gist features from 30 randomly-selected frames from each clip are classified[1], and the resulting probabilities are averaged to provide the clip-level classification probability.

Figure 3 shows the DET curves generated during the cross-validation of training stages.

2.4 MFCCs

MFCCs are one of the most fundamental acoustic features. To capture general audio information of a video, we used MFCCs with a bag-of-words (BoW) model. First, 32-dimensional MFCCs are extracted from a

[1]We also experimented with using the 30 frames nearest to detected keyframes, which produced comparable—but slightly worse—results than randomly-selected frames.



(a) MFCCs (b) ASMs

Figure 4: DET curves for base classifiers per feature type on MED11 Test dataset, for all ten event classes.

video clip at every 10ms with 25ms frame size, and quantized to construct a 1024-sized codebook. Then, we counted codewords with term frequency-inverse document frequency (tf-idf) indexing. After acquiring histograms of codewords, the histogram intersection kernel is used along with SVMs, to form the MFCC base classifier.

The performance of MFCC base classifier for all MED11 event classes are shown in Figure 4(a). It is notable that even though we did clip-level classification with simple SVMs, MFCCs showed competitive performance.

2.5 Acoustic Segment Models (ASMs)

Acoustic segment models (ASMs) are based on the idea that audio can be described as a sequence of smaller acoustic units where each acoustic unit is designed to capture a particular semantic concept. As acoustic utterances could be built up with phones, general audio can be considered as representations of basic acoustic units. General audio, i.e. music, noise, and etc., does not have labels, while speech data has phone labels available. Therefore, to train ASMs, we needed to build initial transcripts. We manually selected 8 representative audio segments for each event class. Then, a hidden Markov model (HMM) training process learns a corresponding ASM for an audio segment. Additionally, we included 39 TIMIT phone models and 5 background noise models.

For decoding, 5-best ASM Viterbi sequences were computed via HMMs where ASM feature vectors were formulated by counting unigrams and bigrams with smooth weighting for bigram counts. For base classifier training, HIK kernels were used within SVM framework.

The performance of ASM base classifier for all MED11 event classes are shown in Figure 4(b). It can be observed that ASMs showed slightly better performance than MFCCs for all event classes. In addition, it has been found that audio features, both MFCCs and ASMs, showed much more discriminative power than video features for particular events where audio provides strong cues. These examples include: event #6, 'Birthday party', and event #14, 'Repairing an appliance'.

3 Late-Fusion Approaches and Experimental Results

Our system incorporates two different late fusion approaches for the TRECVID 2011 MED task.

The details of our primary and secondary runs using all the audio-visual features are described in the following sections as follows:

- `GENIE_MED11_MED11TEST_MEDFull_AutoEAG_p-MFoMaudiovisual_1:` our primary run based on the first fusion approach using all five features is described in Section 3.1.
- `GENIE_MED11_MED11TEST_MEDFull_AutoEAG_c-MLE_1`: our secondary run using alternative fusion approach using all five features is described in Section 3.2.

Interestingly, the performance of two fusion approaches are found to be fairly comparable to each other, showing only marginal differences from each other across all the event types.

The two remaining optional runs with run-ids with the following two run-ids were based on the algorithm used in our primary run where either only audio or only video features were used. The discussion on these two additional runs are omitted for brevity.

- `GENIE_MED11_MED11TEST_MEDFull_AutoEAG_c-MFoMvideo_1`
- `GENIE_MED11_MED11TEST_MEDFull_AutoEAG_c-MFoMaudio_1`

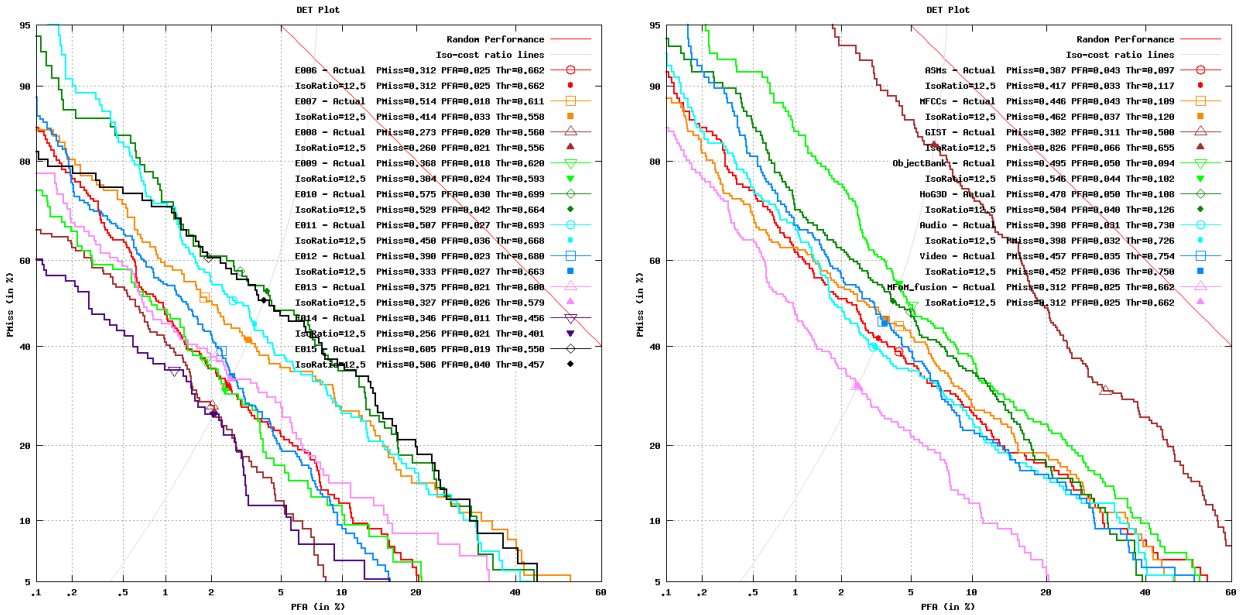
3.1 Maximal Figure of Merit with Model-based Transform

For our primary run, a discriminative score fusion scheme has been used, which is based on the model based transformation (MBT) [9]. The MBT fusion can be regarded as a supervised mapping from the low- or intermediate-level feature space to a high- or semantic-level space. In the MBT fusion, classifiers for each feature are first trained independently. For classification problems with N classes, we learn N individual discriminant functions from the training set using each of K feature. Considering the N discriminant functions as the basis for the space transformation, we can map a given test clip to a new N-dimensional space, where each component implies the similarity between a given sample and a discriminant function, i.e., base classifier outputs. The next step is to learn a new classifier in the model space with mapping from multiple base classifier outputs. To illustrate, if there are K types of features, we can get K number of N-dimensional features in the model space and then concatenate them into a $K \times N$-dimensional feature to describe a sample. In this context, the MBT fusion is categorized as a late fusion method in contrast to an early fusion where multiple features are combined without a semantic-level mapping.

In [2], the advantages of the model-based representation for discriminative score fusion has been presented where it effectively captures semantic concepts. Moreover, it reduces dimensionality and provides a compact and cost-efficient representation, both time and space. Since the MBT-based new feature describes the semantic-level confidence measure and is normalized by the competitive model, it is expected to be more compact with a smaller variance compared to low-level features. Another advantage is that in the MBT fusion, it is easy to fuse multiple distinctive features, e.g. visual, textual, and audio, when they are transformed into a common space. If one feature is more discriminative than others, a fusion classifier will give a heavy weight, and vice versa. The MBT fusion provides a more concrete and fundamental approach than other heuristic fusion methods.

Since 15 event classes are available for the training data of TRECVID MED11 dataset, we trained 15 base-classifiers for each feature, HOG3D, Object Bank, MFCCs, ASMs, and Gist, using support vector machines (SVMs). Although the first 5 events among 15 event classes are not evaluated, we intended to get more discriminative power by including those 5 event classes. Then, all the scores from 15×5 base classifiers are collected and concatenated to construct a new 75-dimensional feature vector in MBT space for discriminative score fusion.

For a fusion classifier, we used maximal figure-of-merit (MFoM) learning scheme, which has shown promising performance especially when a specific performance metric is desired [3, 4]. Let the training data be denoted as $T = \left\{(X, Y) \,|X \in R^D,\, Y \subset C\right\}$, where X is a D-dimensional feature for an instance and Y is the corresponding label. With a binary classification problem, where $C = \{C^+, C^-\}$, MFoM learns a discriminant function, $g(X;\Lambda)$, such that the decision result is positive if $g(X;\Lambda) > 0$ and negative otherwise. Then, the four items in the confusion table, TP (true positive), FP (false positive), TN (true negative), and FN (false negative), can be approximated to continuous forms. For example, the approximation formulas for



(a) (b)

Figure 5: (a) DET curves for discriminative score fusion with MFoM learning scheme on the MED11 test data. (b) DET curves for fusion and base classifiers for the event #6, "Birthday party".

TP and TN are shown below in Eq. 1 and Eq. 2 respectively where $l(X;\Lambda)$ is a class loss function defined as a sigmoid function:

$$TP \approx \sum_{X \in T} (1 - l(X;\Lambda)) \cdot 1\left(X \in C^{+}\right) \quad (1)$$

$$TN \approx \sum_{X \in T} l(X;\Lambda) \cdot 1\left(X \in C^{-}\right) \quad (2)$$

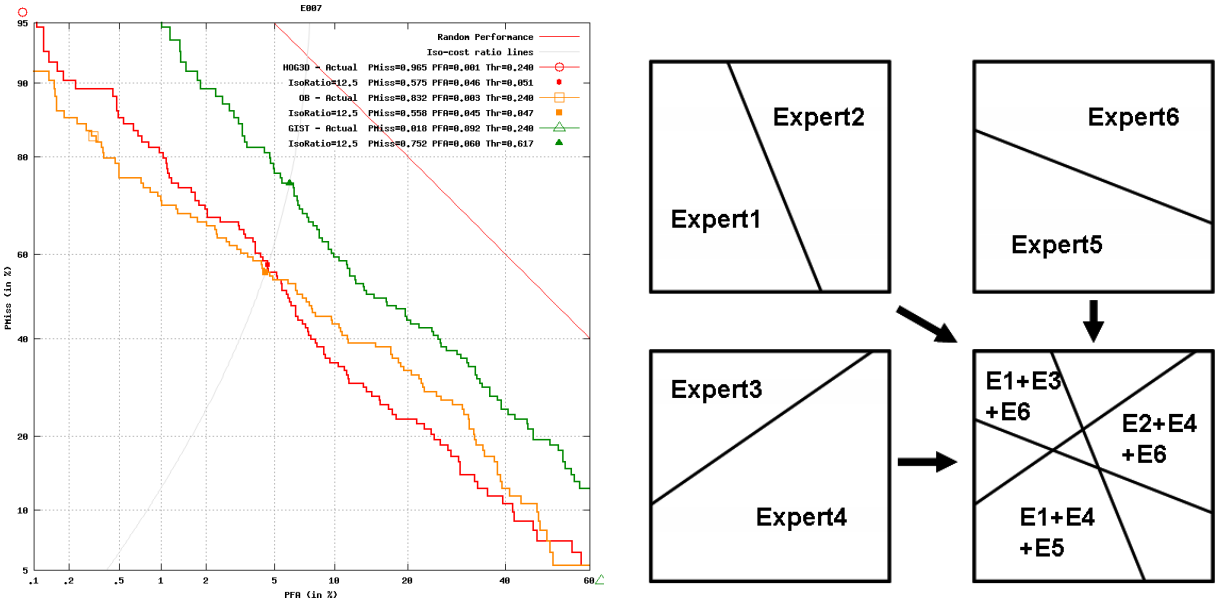
One of the official evaluation metrics for MED11 is normalized detection cost (NDC) [7], which can be considered as a weighted sum of the probability of missed detections and false alarms. In addition, we wanted to consider that a learned threshold can locate a desired target error ratio (TER) among the probability of missed detections (P_{MD}) and the probability of false alarms (P_{FA}). Accordingly, we have incorporated both factors into the MFoM objective function during the training of fusion classifiers.

Since the objective function is continuous and differentiable, we apply a general learning algorithms to estimate the good classifier parameters. Specifically, we used a linear discriminant function for a class discriminant function, so that $g(X;\Lambda) = \omega^T X$.

Figure 5(a) shows the DET curves from discriminative score fusion on the MED11 test data. It is notable that the learned thresholds are aligned around minimum NDC points and the 12.5 iso-line. We have also observed that our discriminative score fusion shows improvements for all the event classes compared to base classifiers. In Figure 5(b), we can compare the DET curves of base classifiers, fusion with only audio, fusion with only video, and fusion with audio and video for E006 'Birthday party'. Especially, we could see that the performance gain from fusion is significant around the 12.5 iso-line.

3.2 Mixtures of Local Experts

Our second score fusion approach was based on a newly-designed algorithm dubbed the Mixture of Local Experts (MLE). When we compared the performance of the GENIE base classifiers by plotting their DET curves for a given event on the same axes, as in Figure 6(a), two categories of base classifier pairs arise. In the first category, we have base classifier pairs where one consistently outperforms (i.e., has a DET curve that is lower than) the other. In our example figure, Object Bank and HOG3D classifiers both consistently



(a) (b)

Figure 6: (a) DET curves for Gist, HOG3D, and Object Bank base classifiers. (b) A notional mixture of local experts, where a two-dimensional score space is partitioned with three hyperplanes.

out-perform the Gist classifier. For such a relationship, we can learn a linear combination of their output probabilities where the better classifier's scores are given higher weight. This approach does not work well, however, when the DET curves for two classifiers cross, as the Object Bank and HOG3D curves do in our example figure. These crossings show that the relative performance of the classifiers depends on the score that each produces; in our example, the Object Bank classifier out-performs HOG3D at high probabilities and HOG3D outperforms Object Bank at low probabilities.

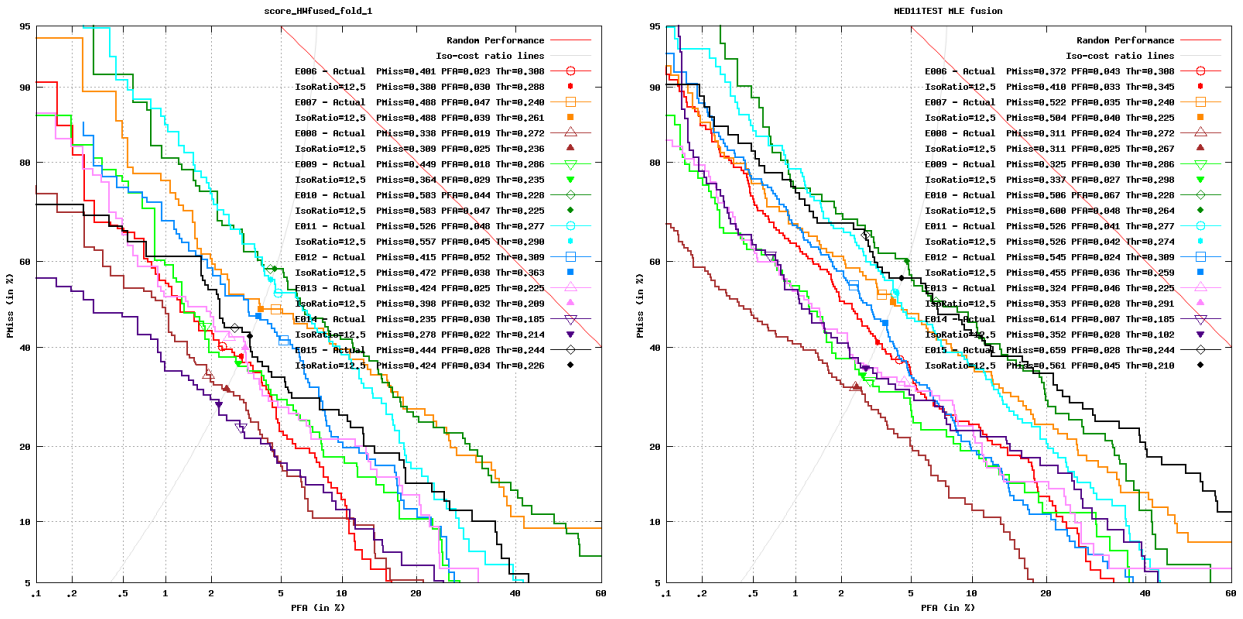
For N base classifiers, we consider each clip as a point in an N-dimensional score space. In order to capture score-based variations in the relative performances of the base classifiers, our MLE fusion partitions this space and computes local weightings independently for the different partitions. In order to search over the infinite number of partitions, we employ a greedy search strategy for partitioning. We iterate by considering new partitions and keeping those that improve the fusion performance on the training data.

Figure 7(a) shows the DET curves produced by the MLE algorithm during training. Figure 7(b) shows the DET curves on the MED11 test data, as computed by NIST. The performance is comparable between the two, and the relative ordering of the events is similar. Note that, due to time constraints imposed by the MED11 deadline, our MLE algorithm did not include an analytic method to determine thresholds. Instead, we set our submission thresholds based on the average probability value, over the five folds of testing, at the intersection of the 12.5 iso-line. It can be observed that the estimated thresholds based on this simple scheme is fairly effective, reappearing close to the iso-lines on MED11TEST data. Note that these do not necessarily minimize Normalized Detection Cost (NDC).

4 Classification based on metadata

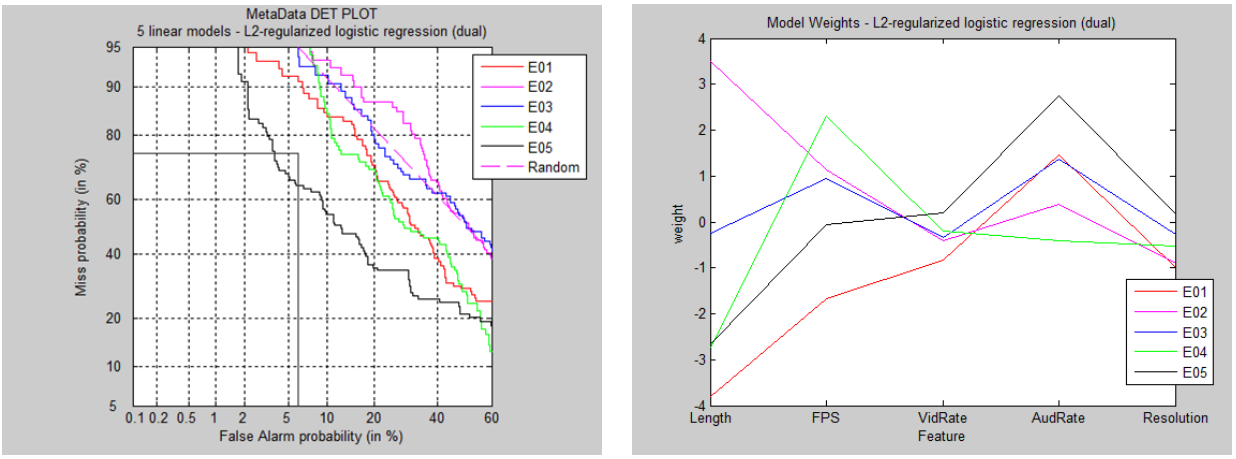
As a side experiment, we also looked at classifying the clips into events without any audio or video data per se, by simply using some of the metadata: clip length, frame rate, video bitrate, audio bitrare, and frame size. DET curves for training events 1–5 are shown in Figure 8(a). We found that 3 of the 5 event classifiers performed better than random, and that one of them (event 5: woodworking) shows fairly competitive performance, achieving 75% PMiss at 6% PFA!

An in-depth analysis of the model weights (visualized in Figure 8(b)) shows that the correlation is not semantically meaningful. For example, one could argue that wedding videos should long and high quality



(a) (b)

Figure 7: (a) DET curves for MLE-based fusion, produced during training for MED11. (b) Actual DET curves on the MED11 test data.



(a) (b)

Figure 8: (a) DET curve for classification with metadata alone. (b) The SVM model weights for the different metadata dimensions.

(high bitrate), but that is not supported by the model weights.

This points to the possibility that there is some correlation in the dataset, perhaps in the collection methodology, or inherent in the nature of Internet videos.

5 Conclusion

We have presented our GENIE system which incorporated two novel fusion techniques built upon five base classifiers. Considering that standard SVM approaches are used as base classifiers, and almost no event-specific adaptations were made, our results can be thought to constitute a very strong baseline.

Interestingly, no major or systematic performance gap has been observed between the two fusion approaches we employed. Such results suggest that additional features, more detailed segment-level analysis (instead of clip-level analysis), and interactive learning where users aid the system to identify more meaningful subset of event kit examplars are the next research venues to explore for improving MED performance in the future.

6 Acknowledgments

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] 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.

[2] Sheng Gao, De-Hong Wang, and Chin-Hui Lee. Automatic image annotation through multi-topic text categorization. In *ICASSP*, 2006.

[3] Sheng Gao, Wen Wu, Chin-Hui Lee, and Tat-Seng Chua. A mfom learning approach to robust multiclass multi-label text categorization. In *ICML*, 2004.

[4] Ilseo Kim and Chin-Hui Lee. Optimization of average precision with maximal figure-of-merit learning. In *MLSP*, 2011.

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

[6] 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.

[7] NIST. 2011 TRECVID Multimedia Event Detection Evaluation Plan Version 3.0. http://www.nist.gov/itl/iad/mig/upload/MED11-EvalPlan-V03-20110801a.pdf.

[8] 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.

[9] De-Hong Wang, Sheng Gao, Qi Tian, and Wing-Kin Sung. Discriminative fusion approach for automatic image annotation. In *MMSP*, 2005.