Team SRI-Sarnoff's AURORA System @ TRECVID 2011

Hui Cheng†, Amir Tamrakar†, Saad Ali†, Qian Yu†, Omar Javed†, Jingen Liu†, Ajay Divakaran†, Harpreet S. Sawhney†, Alex Hauptmann♦, Mubarak Shah♠, Subhabrata Bhattacharya♠, Michael Witbrock♡, Jon Curtis♡, Gerald Friedland✧, Robert Mertens✧, Trevor Darrell✧, R. Manmatha❋, James Allan❋

† SRI-International Sarnoff, Vision Technologies Lab, 201 Washington Road, Princeton NJ 08540
♦ School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213
♠ Computer Visions Lab, University of Central Florida, Orlando, FL 32816
♡ Cycorp Inc., Austin, TX 78731
✧ International Computer Science Institute, University of California–Berkeley, Berkeley CA 94704
❋ University of Massachusetts-Amherst, Amherst, MA 01003

Abstract

In this paper, we present results from the experimental evaluation for the TRECVID 2011 MED11 (Multimedia Event Detection) task as a part of Team SRI-Sarnoff's AURORA system being developed under the IARPA ALADDIN Program. Our approach employs two classes of content descriptions for describing videos depicting diverse events: (1) Low level features and their aggregates, and (2) Semantic concepts that capture scenes, objects and atomic actions that are local in space-time. In this presentation we summarize our system design and the content descriptions used. We also present four MED11 experiments that we submitted, discuss the results and lessons learned.

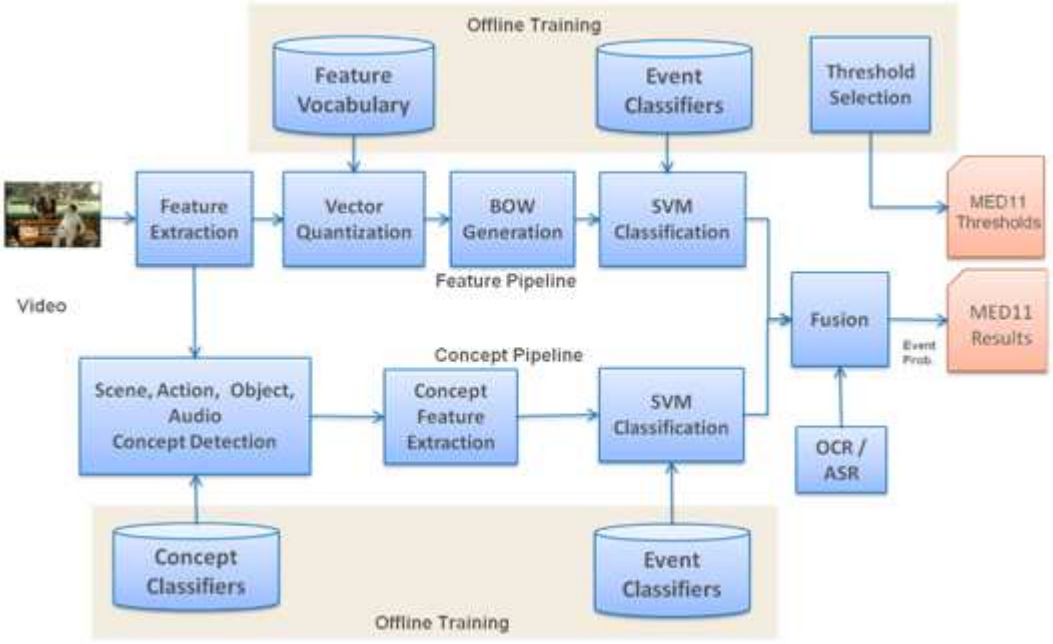

Figure 1 The overall approach for multimedia event detection.

1 Introduction

Team SRI-Sarnoff participated in the Multimedia Event Detection (MED11) task in the 2011 TRECVID evaluation program. The goal of MED11 is to "promote content-based analysis of and retrieval from digital video via open, metrics-based evaluation". The stated problem description for the MED task is as follows: *"Given a collection of test videos and a list of test events, indicate whether each of the test events is present anywhere in each of the test videos and give the strength of evidence for each such judgment."*

This paper documents SRI-Sarnoff's approach, experiences, experimental results, observations and lessons learned.

2 Approach to Multimedia Event Detection

We evaluated two classes of approaches for event representation and retrieval as shown in Figure 1. The feature-based approach is a direct supervised training approach to event detection based on a number of low-level audio and visual features and their aggregates. The second is a concept-based approach striving for semantic content-based description of events in video. We describe these approaches in more detail in the following sections. We submitted separate results from the two approaches as well as a third result that was produced by fusing the results from the two approaches in a late-fusion step. A final step of threshold selection for choosing an optimal operating point for the system is described later in the Experiments section.

2.1 Audio-Visual Feature-Based Event Detection

We selected a variety of features to capture various aspects of an event (e.g. scene, motion, audio) listed in Table 1. These features were computed either on single frames or on spatio-temporal windows of frames (XYT-cubes) throughout a given video clip. The event unfolding in a video clip is represented as an aggregate feature -- the histogram of "words" corresponding to each feature type computed over the entire video clip. This is popularly known as a "Bag-of-Words" (BoW) representation.

In order to compute BoW descriptors for each feature type, feature specific vocabularies are first learned using k-means clustering of raw features. For static features (e.g. SIFT, GIST) we generated a vocabulary of 1000 words while for motion features (STIP, Trajectory) we used a vocabulary of 10000 words. For MFCC features a random forest [7] based clustering approach is used to generate the vocabulary. Once the features in a video are quantized using the respective vocabularies, a BoW is computed per feature. Event models are trained using SVM [8] with intersection kernel. For exploring combination of features, BoWs were concatenated before computing the kernel matrix for SVM.

Table 1 Low-level Features and what they attempt to capture.

Feature	Description
Hessian-Affine Interest Point [1] + SIFT (Scale Invariant Feature Transform) [2]	Captures local image gradient structure around corner features in the image
Hessian-Affine Interest Point + Color SIFT [3]	Captures local image gradient structure in RGB color space around corner features in the image
GIST [4]	A global image descriptor that captures gradient structure over and entire image at various scales and orientations
MOSIFT (Motion SIFT) [9]	Image gradient around sparsely tracked SIFT points
STIP [5]:	Spatio-temporal Interest Points defined by gradient and optical flow structure in a fixed window around 3D corner features
Trajectory Motion Boundary Histogram [6]:	Trajectory centric local motion statistics
Trajectory Motion Boundary Histogram– Histogram of Gradients [6]:	Trajectory centric local image gradient statistics
MFCC (Mel Frequency Cepstrum Coefficients) :	Short-term power spectrum of a sound

2.2 Concept-Based Event Detection

One of the challenges for event recognition is to bridge the semantic gap between low-level features and high-level events. This is precisely what semantic concepts trained on the low-level features are designed to accomplish. There are a variety of reasons to represent events in terms of semantic concept features. Concepts are directly connected to the Event Kit Descriptions. Concept-based event representation enables the recognition system to integrate multi-modality information such as human knowledge and Internet resources. Furthermore, thanks to the semantic meaning of concepts, the concept-based event representation potentially has better generalization capability, which is significantly important for event recognition, especially when only a few training examples are available.

We defined four classes of concepts for MED11: actions, scenes, objects and audio concepts (See Table 2). For visual action concepts (atomic *and localized motion and appearance patterns*), we selected 81 mostly human action centric concepts based on Event Kit descriptions and video examples. We annotated more than 40 examples for each concept. In addition, we defined 18 scene concepts including indoor scenes (e.g., kitchen, living room, church, etc.) and outdoor scenes (e.g., street, highway, river, etc.). We also used about 20 audio concepts.

From the video clips (shots), we extract various low-level features to represent the concepts. Action concepts include three types of complementary low-level features: dense trajectory based HOG, dense trajectory based MBH (motion boundary histogram), as well as STIP. For scene concepts we extract SIFT and GIST features, while HOG based deformable model is applied for object detection. To detect audio concepts, we compute the MFCC features. Bag of words model is used to train the SVM-based action concept detection and Random-Forest-Based audio concept detection. The trained concept detectors are applied to moving XYT windows in an unknown video.

Table 2 Concept Examples. Only a small subset of concepts used is included.

Action Concept (81)
Animal approaching
Animal eating
Open door
People dancing
People marching
Person bending
Person blowing candles
Person carving
Person casting
Person clapping
Person climbing
Person cutting
Person cutting fabric
Person dancing
Person drilling
Person falling
Person flipping
Person hammering

Audio Concepts (20)
animal sound chewing eating
board hitting surface
cheering
clapping
crowd noise
drilling
engine noise from vehicle
hammering
laughing
metallic clanking clicking popping
music
planing
power tool whine
rolling
sawing
scraping
sewing machine sound
speech

	Scene Concepts (18)		
Close-up	Home	Urban	Natural
Animal	Kitchen	City Street	Lake/Pond
Food	Shed / Garage	City Square	Ski Slope
Hands		City Park / Garden	Sea
Wheel		Residential Area	Woods/ Forest
Machine		Church	
Text Only			
Unknown			

Object Concepts (4)
Face frontal
Face profile
person
car

Features derived from Concepts for Event Detection

Each concept detector is applied to overlapping XYT windows exhaustively within a given video clip. As a result, for each XYT window, a vector of *N* concept detector scores is generated -- one score for each of the *N* concepts. We derive aggregated features from these *N*-vectors to represent the event present in the video. We explore a variety of aggregate features: Max_Avg_Std, Concept Histogram, Concept Co-occurrence Matrix, and Max Outer Product. These are described in Table 3 below.

Table 3 Description of Features derived from concepts

Feature	Description
Max Avg Std	• Calculate Max, Average, and Std. Deviation of the detection scores • Captures the shape of single Gaussian model of concept distribution • Feature dimension: number of concepts x 3
Histogram Vector	• Compute the frequency of occurrence of a concept in a video • Capture the direct occurrence (*first order*) of concepts • Feature dimension: number of concepts
Co-occurrence Mat	• Compute the chance of happening between every pair of concepts • Captures the overall concept co-occurrence *(second order)* • Feature dimension: number of concepts x number of concepts
Max Outer Product	• Take the Max concept detection confidence values of a video and make an outer product from these • Captures the co-occurrence of the most confident concepts • Feature dimension: number of concepts x number of concepts / 2

3 Experiments

3.1 Training/Testing Methodology

We adopted the Support Vector Machine (SVM) as our basic classifier and use intersection kernels for all histogram-based features and RBF (Radial Basis Function) kernels for concept-based features. Cross validation was used to determine the SVM parameters before the actual training. We applied L1 normalization to histogram-based features. The videos from the event kits were used as positive samples and all non-event videos were used as negative samples to train a binary classifier for each event independently. Each classifier output a probability of detection as a score. LibSVM [8] is used as the SVM solver.

Standard Training/Testing Evaluation Folds: In order to prepare for MED11 evaluation, we separated the Event Collection (EC) and Development Collection (DEVT) data into a set of standard evaluation folds to avoid the differences caused by dataset variations. Since DEVT only contains positive clips for events 1-5, we moved 40% of data for events 6-15 from EC into DEVT. We call this dataset "*Mixed-DEVT*". All of our experiments, from exploring good features to fusing features from different modalities, are based on these standard evaluation folds.

In all event classification evaluation experiments described below, we use the Missed Detection (MD) rate at 6% False Alarm (FA) rate as the metric for comparison.

3.2 Feature-Based Experiments

Experiments were performed to explore the best mix of low-level image and video features described earlier. In terms of feature combinations, we discovered that if features are chosen judiciously then the more the features the better the performance. In a feature by feature comparison, the trajectory centric features consistently performed better than any other feature type. Also, combining features that are computed on different underlying key-frames gives improvement – i.e. higher rate sampling of key-frames within a shot is useful. Table 4 shows the MD rates at 6% FA on the mixed-DEVT (event kit+ development set) data set.

Table 4 Results on Mixed-DEVT for each feature and different feature combinations

Event	GIST	STIP	SIFT	MFCC	DTF MBH	CMU MOSIFT	STIP-SIFT-GIST-MFCC	STIP-SIFT-GIST-MFCC –DTF_MBH-DTF_HOG + CMU_SIFT-CMU_MOSIFT-CMU-CSIFT

Attempting a board trick	0.61	0.44	0.55	0.63	0.22	0.56	0.33	0.18
Feeding an animal	0.8	0.84	0.71	0.81	0.66	0.74	0.65	0.58
Landing a fish	0.55	0.42	0.56	0.8	0.23	0.35	0.29	0.16
Wedding ceremony	0.44	0.28	0.31	0.51	0.22	0.33	0.22	0.17
Working on a woodworking project	0.68	0.57	0.49	0.54	0.51	0.56	0.29	0.21
Birthday party	0.71	0.53	0.7	0.49	0.56	0.57	0.34	0.3
Changing a vehicle tire	0.77	0.72	0.63	0.87	0.79	0.62	0.57	0.43
Flash mob gathering	0.28	0.2	0.21	0.44	0.35	0.31	0.15	0.12
Getting a vehicle unstuck	0.76	0.4	0.51	0.57	0.54	0.49	0.34	0.24
Grooming an animal	0.87	0.71	0.68	0.87	0.71	0.59	0.59	0.48
Making a sandwich	0.74	0.55	0.78	0.8	0.61	0.67	0.51	0.43
Parade	0.54	0.5	0.41	0.76	0.36	0.48	0.26	0.2
Parkour	0.66	0.25	0.43	0.85	0.31	0.41	0.25	0.15
Repairing an appliance	0.5	0.3	0.39	0.34	0.48	0.31	0.23	0.19
Working on a sewing project	0.82	0.5	0.65	0.64	0.61	0.56	0.42	0.34
AVERAGE	**0.65**	**0.48**	**0.53**	**0.66**	**0.48**	**0.5**	**0.36**	**0.28**

3.3 Concept-Based Experiments

We first constructed visual vocabularies of size 10,000 for each of the low-level features (including DTF-HOG, DTF-MBH, and STIP). The action concept detectors were then trained on the BOW descriptors compiled using these vocabularies for the annotated video clips from EC data set. We then performed cross-validation for action concept classification to verify the complementary properties of three types of low-level features. Table 5 shows the concept classification results for a few of the concepts used in terms of MD rates at 6% FA rates. As we can see, the combination of DTF MBH-HOG performs best for most concepts, while MBH and HOG are individually competitive. Both of them perform better than STIP features.

Table 5 Concept Classification Performance on Annotated Video Clips

	Spatio-Temporal	Trajectory Based Descriptors		
Action Concept	**STIP**	**MBH**	**HOG**	**MBH-HOG**
Animal approaching	0.6778	0.6667	0.5667	0.5000
Animal eating	0.6914	0.3219	0.2286	0.1981
Open door	0.3810	0.1667	0.2333	0.1333
People dancing	0.1809	0.0889	0.1289	0.0711
People marching	0.2039	0.1867	0.1733	0.1467
Person bending	0.7442	0.7000	0.6762	0.5762
Person blowing candles	0.7143	0.6963	0.5037	0.4074

For event detection, we tested our approach using various concept features on the mixed-DEVT data set defined above. The MD rates at 6% FA rates are reported in Table 6. It is worth noting that we separated Events 1-5 from Events 6-15. This is because the testing videos of Events 1-5 are from DEVT, which does not have any overlap with the videos used for training concept detectors. As for the rest of the events, the testing videos may have come from EC data set, which had previously been used for training our action concept detectors. Therefore, while the results on Events 1-5 exhibit the actual event detection performance, the results on Events 6-15 may be biased. As we can observe from the table, the concept features are mostly complementary although some features perform

relatively better. Overall, we obtain the best performance by fusing the results of all types of concept features. Following this observation, we adopted all the concept features for testing on MED11 test videos.

Table 6 Event Detection Performance using Concept-Based Features on the Mixed-DEVT Data Set

Event	Histogram	Co-occurrence	Max-outer-product	Max-Avg-Std	Fusion
Attempting a board trick	0.171	0.173	0.199	0.214	0.157
Feeding an animal	0.631	0.621	0.688	0.716	0.621
Landing a fish	0.250	0.202	0.276	0.287	0.212
Wedding ceremony	0.232	0.209	0.283	0.289	0.194
Working on a woodworking	0.514	0.464	0.437	0.417	0.380
Birthday party	0.429	0.400	0.191	0.181	0.226
Changing a vehicle tire	0.262	0.216	0.131	0.084	0.120
Flash mob gathering	0.274	0.170	0.134	0.137	0.141
Getting a vehicle unstuck	0.172	0.189	0.157	0.134	0.149
Grooming an animal	0.454	0.429	0.405	0.384	0.350
Making a sandwich	0.227	0.216	0.088	0.106	0.110
Parade	0.284	0.250	0.171	0.155	0.139
Parkour	0.118	0.140	0.122	0.113	0.091
Repairing an appliance	0.296	0.252	0.288	0.226	0.186
Working on a sewing project	0.273	0.283	0.183	0.179	0.156
AVERAGE(1-15)	0.306	0.281	0.250	0.242	0.215
AVERAGE (6-15)	0.279	0.254	0.187	0.170	0.167
AVERAGE (1-5)	0.360	0.334	0.377	0.385	0.313

3.4 Threshold Determination Approach

The MED11 task required all systems to select an overall operating point. We ran several experiments on the development dataset (DEVT) for gauging the sensitivity of the missed detection (MD) rates and False alarm (FA) rates to the selected thresholds. We found (Figure 2) that across different partitions of the training data, the MD rate as a function of threshold was very unstable where as the FA rate as a function of the threshold was remarkably stable. This observation therefore led us to pick the strategy of selecting thresholds based solely on the FA rates.

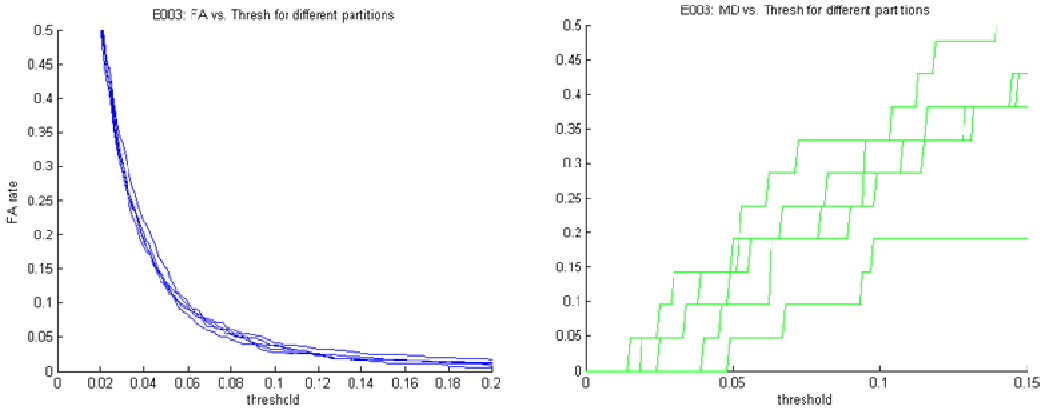

Figure 2 [Left] FA vs. threshold curves and [Right] MD vs. threshold curves for different partitions of the data.

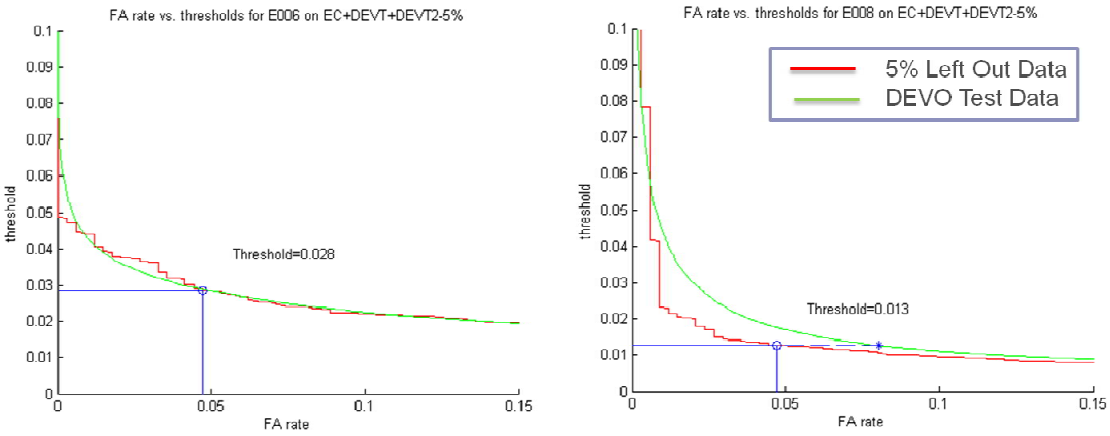

Figure 3 FA-vs-threshold curves from the 5% left out data compared to the actual curves obtained from the test data.

We also found that the sensitivity to thresholds do not match across events. This meant that we would not be able to estimate thresholds from the training events in the DEVT dataset. We, therefore, left behind 5% of the negative training data randomly and used the rest of the data to train the classifiers. The left behind data was then passed through the event classifier to obtain the FA-vs-threshold curve for each event. From this curve we picked thresholds that would give us an operating point at 5% FA rate.

Figure 3 shows the efficacy of that approach to the MED11 task. Having received the ground truth for the MED11 test data, we computed the FA-vs-threshold curves on the entire MED11 test set (DEVO). The green curve shows this relationship for the test data and the red curve shows the estimated one based on the training data. Some of the estimated curves were very representative of the larger dataset and others were somewhat off. We have not as of yet determined the reason for this variation.

4 MED11 Results and Discussion

All the computations reported in this paper were performed on the SRI-Sarnoff AURORA system. This system comprises of a number of servers with web interfaces for browsing the datasets, annotating the training data as well as managing the experiments run over a distributed computational pipeline. The computational pipeline currently consists of 120 AMD Opteron nodes with 5GB RAM per node as well as a number of nVidia Tesla M2050 GPUs and is based on the Apache UIMA (Unstructured Information Management Architecture) which is essentially a highly configurable filter graph like architecture that allows for process distribution across multiple nodes.

We submitted four sets of results to TRECVID MED11:

1. **Primary Run (SRI-AURORA_MED11_MED11TEST_MEDFull_AutoEAG_p-LateFusion_1):**
   Results of the feature-based approach (described in Section 2.1) where separate classifiers were trained on each low-level feature and their outputs combined via a late fusion step.

2. **Contrastive Run 1 (SRI-AURORA_MED11_MED11TEST_MEDFull_AutoEAG_c-ConceptsLateFusion_1):**
   Results of the concept-based approach (described in Section 2.2) where separate classifiers trained on each concept-based feature were combined in a late fusion step.

3. **Contrastive Run 2 (SRI-AURORA_MED11_MED11TEST_MEDFull_AutoEAG_c-LLAndConceptsLateFusion_1):**
   This submission combined the results from the feature-based approach in submission 1 with the concept-based approach in submission 2 in a combined late fusion step.

4. **Contrastive Run 3 (SRI-AURORA_MED11_MED11TEST_MEDFull_AutoEAG_c-EarlyFusion_1):**
   This submission contained the results of our feature-based approach similar to our primary submission but

with the features combined in an early fusion step.

In Table 7 below, we report the average performance of each of these submissions at the automatically selected operating points as well as the computational times (wall clock times) on our system. The complete set of DET curves on all test events for all four submissions are shown in Figures 4, 5 and 6.

Table 7 MED 11 Results.

Run type	MD (%)	FA (%)	Feature Extraction (hrs)	Training Time (hrs)	Testing Time (hrs)
Feature-based Late Fusion (p)	**22.2**	**5.6**	620	10	6
Concept-based Late Fusion (c1)	**32.1**	**5.4**	850	2	2
Fusion of Features-based and Concept-based (c2)	**26.4**	**5.1**	850	12	8
Feature-based Early Fusion (c3)	**33.8**	**5.3**	620	28	12

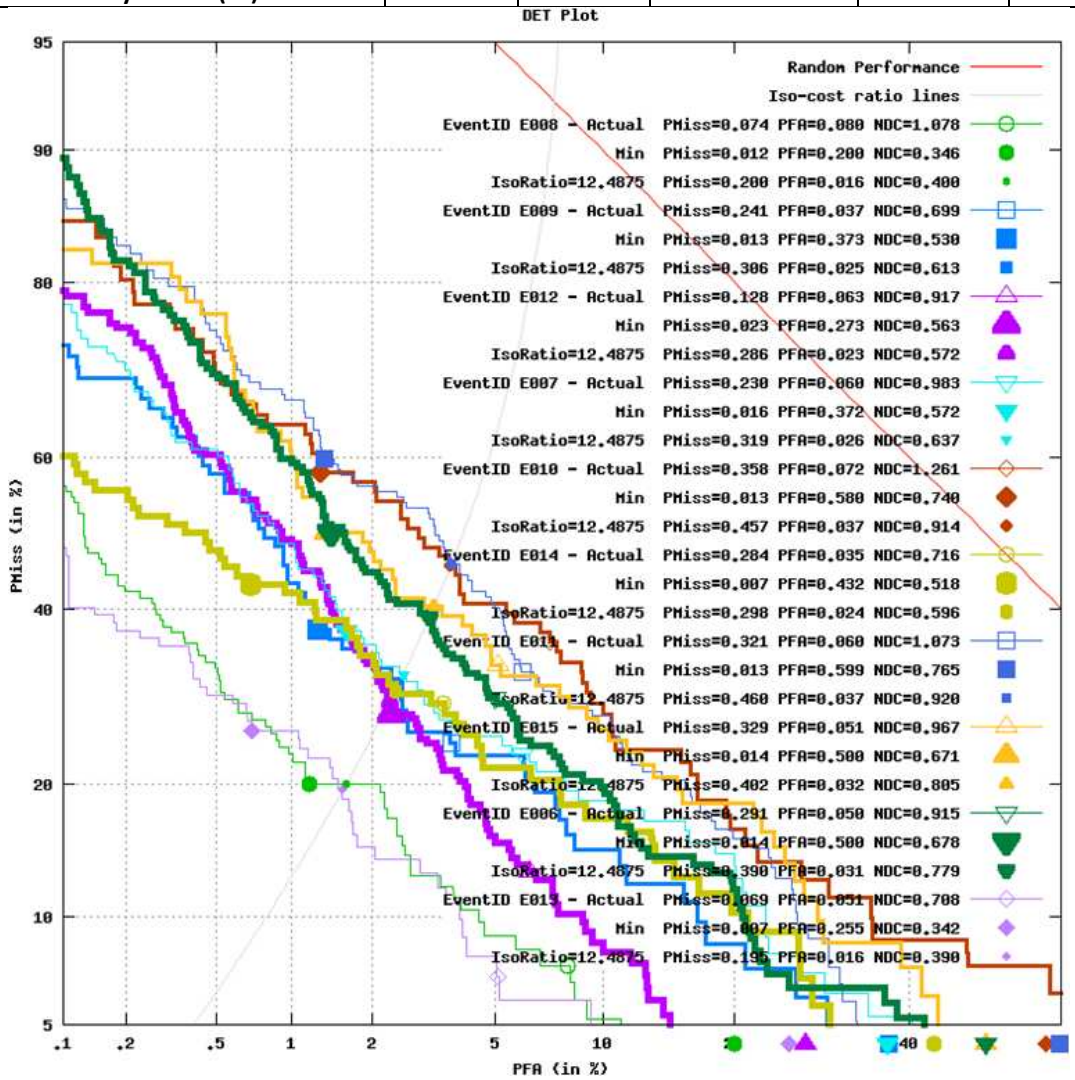

Figure 4 The DET curves on MED11 Test Data (DEVO) for Late Fusion of Features

Discussion

Our primary goal with this effort was to setup a system for tackling the content based multimedia retrieval problem and obtaining a baseline performance for the event detection task using low-level features as well as semantic features, which we were able to achieve in a short span of time as evidenced by our TRECVID MED11 submissions.

From our experiments, we observed that the BOWs descriptions of events derived from low-level features which are essentially histograms of various feature vocabularies compiled over the *entire* length of the video clip works surprisingly well for event detection, as depicted by the results of our Primary Run. Overall, we observed that with "judiciously chosen features", the more features we include the better the performance.

The DET performance curves, on the MED11 test data, broadly clustered into three groups: (1) parkour, flash mob; (2) getting a vehicle unstuck, repairing an appliance, parade, making a sandwich; and (3) changing a tire, birthday party, sewing, and grooming an animal. While we can explain the good performance on the first cluster of events as being due to our motion-feature heavy representation, similar explanations are harder to come by for the other two classes of events. We are currently working on methods for analyzing the impact of the various features on the overall performance.

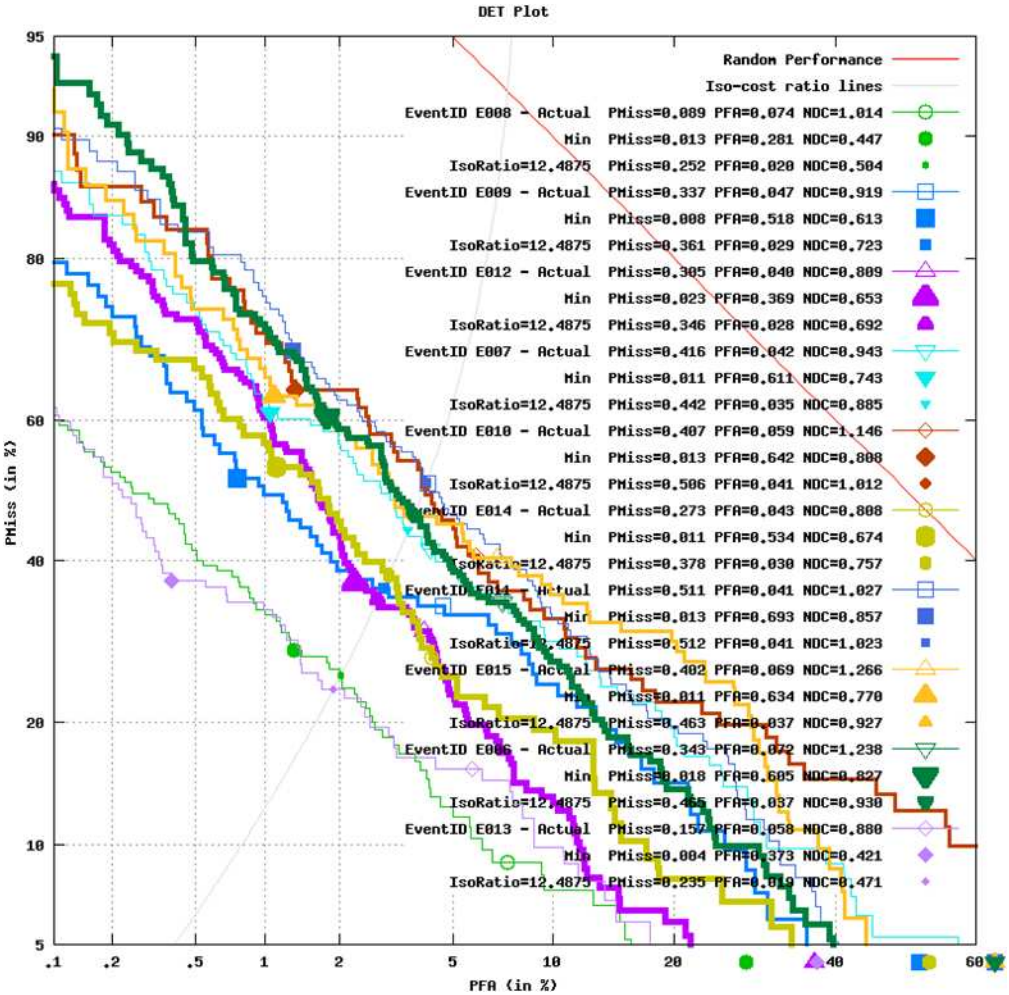

Figure 5 The DET curves on MED11 Test Data (DEVO) for the Concept-only run.

Remarkably, despite the fact that our concept detectors are not well matured (mostly based on global descriptions

of XYT cubes) *and* the fact that we only used a small number of them, the achieved performance is quite reasonable and rather reassuring because in the future with ad hoc events, successful approaches will need to be based on semantically meaningful concepts rather than feature based approaches.

Not surprisingly, the concept-based run also mimics the performance trends of the feature-based run. Preliminary analysis of these results suggests that this approach is performing well for events with high level of actions, moderately with medium levels of actions, and worst with events with low action levels. We are continuing to evaluate the impact of each concept on the detection of events for each specific test video. Along the same vein, we are also analyzing the sensitivity of event detection performance to the choice of the concept bases.

Future work includes better description of videos in terms of low-level features, using larger collections of concepts, obtaining higher accuracy in concept detection by exploiting both spatial and temporal relationships between concepts, composition of events from concepts, automatic selection of high-value concepts from high-level knowledge, and assessing the impact of individual concepts on the event detection performance.

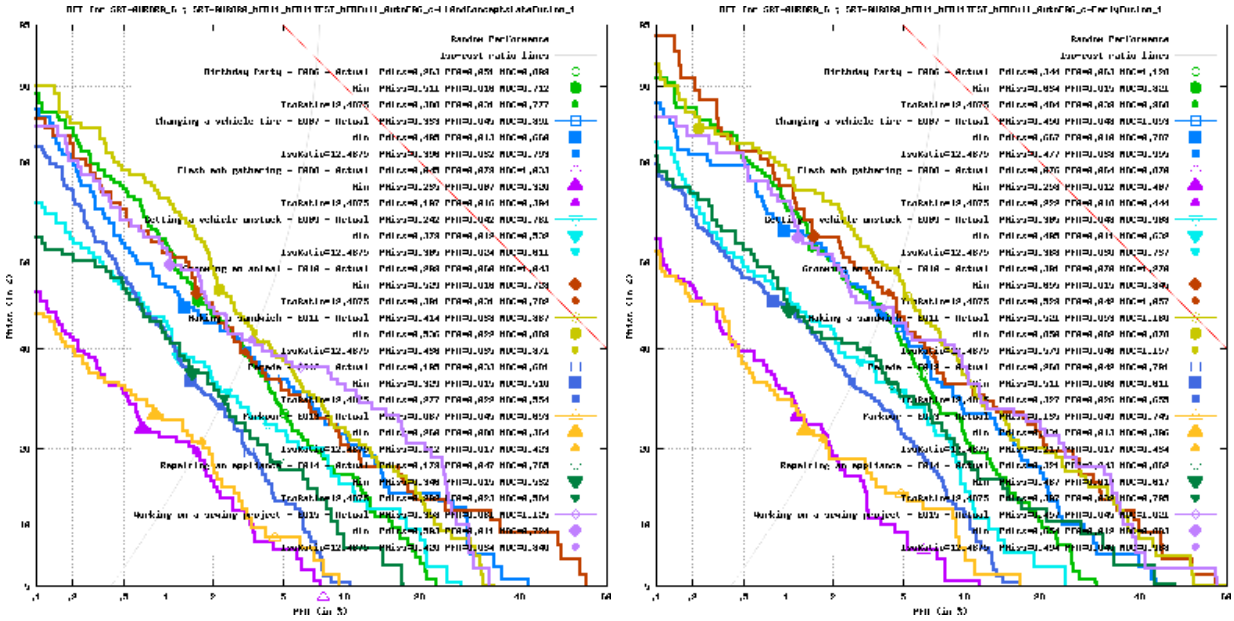

Figure 6. The DET curves on MED11 Test Data (DEVO) for the Low-level + Concepts run and the Low-level Early Fusion run.

References

1. C. Schmid, K. Mikolajczyk, Scale and Affine invariant interest point detectors. IJCV, pp. 63-86. 2004
2. D. Lowe, Distinctive image features from scale invariant keypoints. IJCV, pp. 91-110. 2004
3. J. M.Geusebroek, G. J. Burghouts Performance evaluation of local colour invariants. CVIU, Vol. 113, pp. 48-62. 2009
4. A. Torralba and A. Oliva Modeling the shape of the scene: a holistic representation of the spatial envelope. 2001, IJCV.
5. Ivan Laptev, Tony Lindeberg: Space-time Interest Points. ICCV 2003: 432-439.
6. H. Wang, A. Kläser, C. Schmid and C.-L. Liu, Action Recognition by Dense Trajectories, CVPR2011
7. L. Breiman, Random Forests, Machine Learning , pp. 5-32. 2001
8. C.-C Chang and C.-J. Lin LIBSVM : a library for support vector machines. ACM T-IST, pp. 1-27.2011
9. M. Chen and A. Hauptmann MoSIFT: Reocgnizing Human Actions in Surveillance Videos. CMU-CS-09-161, 2009.