Columbia-UCF TRECVID2010 Multimedia Event Detection: Combining Multiple Modalities, Contextual Concepts, and Temporal Matching

Yu-Gang Jiang[1], Xiaohong Zeng[1], Guangnan Ye[1], Subhabrata Bhattacharya[2],
Dan Ellis[1], Mubarak Shah[2], Shih-Fu Chang[1]

[1] Department of Electrical Engineering, Columbia University
[2] Department of EECS, University of Central Florida

Summary

TRECVID Multimedia Event Detection offers an interesting but very challenging task in detecting high-level complex events (Figure 1) in user-generated videos. In this paper, we will present an overview and comparative analysis of our results, which achieved top performance among all 45 submissions in TRECVID 2010.

Our aim is to answer the following questions. What kind of feature is more effective for multimedia event detection? Are features from different feature modalities (e.g., audio and visual) complementary for event detection? Can we benefit from generic concept detection of background scenes, human actions, and audio concepts? Are sequence matching and event-specific object detectors critical?

Our findings indicate that spatial-temporal feature is very effective for event detection, and it’s also very complementary to other features such as static SIFT and audio features. As a result, our baseline run combining these three features already achieves very impressive results, with a mean minimal normalized cost (MNC) of 0.586. Incorporating the generic concept detectors using a graph diffusion algorithm provides marginal gains (mean MNC 0.579). Sequence matching with Earth Mover’s Distance (EMD) further improves the results (mean MNC 0.565). The event-specific detector (“batter”), however, didn’t prove useful from our current re-ranking tests. We conclude that it is important to combine strong complementary features from multiple modalities for multimedia event detection, and cross-frame matching is helpful in coping with temporal order variation. Leveraging contextual concept detectors and foreground activities remains a very attractive direction requiring further research.

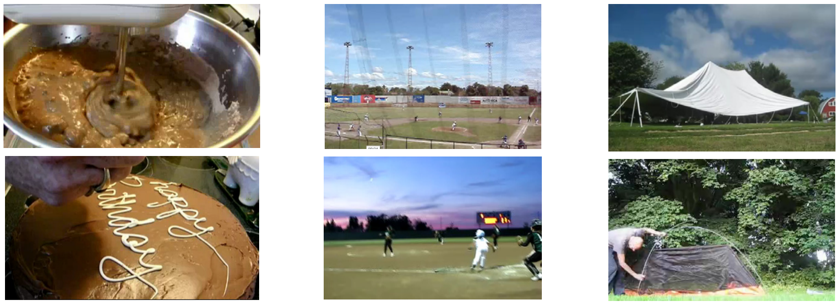
Description of Submitted Runs

Run6	Baseline – average fusion of 3 SVM classification results for each event using 3 feature modalities: 1) spatial-temporal interest points, 2) SIFT, and 3) bag of MFCC audio words.
Run5	Contextual diffusion of Run6 using scene and audio concept detectors.
Run4	Contextual diffusion of Run6 using scene, audio and human action concept detectors.
Run3	Linear fusion of Run6 with a SVM classification result using temporal EMD kernel.
Run2	Contextual diffusion of Run3 using scene, audio and human action concept detectors.
Run1	Re-ranking of event “batting in run” from Run2 using an event-specific “batter” detector.

1. Introduction

Automatic detection of complex events in unconstrained videos has great potential for many applications, such as web video indexing, consumer content management, and open-source intelligence analysis. It is a challenging task due to large content variation and uncontrolled capturing conditions (cf. Figure 1). However, due to the explosive growth of the user generated videos on the Internet, this problem has received a lot of interest from the research community and funding programs [1, 2, 3, 13, 14].

In TRECVID 2010, a new Multimedia Event Detection (MED) task is established to advance research and development in this area. The aim of MED is to develop systems that can automatically find video clips containing any event of interest, assuming only a limited amount of training exemplars are given. Figure 1 displays a few example video frames of the three events evaluated in TRECVID MED 2010.



Making a cake *Batting a run in* *Assembling a shelter*

Figure 1: Example frames of the three events evaluated in TRECVID MED 2010. Content of the same event class can be very different.

In our MED 2010 system, we explored several interesting and important issues including video feature representation, temporal matching, event contexts, and reranking by event-specific object detector. In the following we discuss each of the components in detail.

2. Feature Representation

Feature representation is critical for video content understanding. In TRECVID 2010, we explore three feature modalities for multimedia event detection.

Static SIFT feature

We adopt two sparse keypoint detectors: Difference of Gaussian (DoG) [4] and Hessian Affine [5]. Since the two detectors extract keypoints of different properties, we expect that they are complementary. Using multiple keypoint detectors is also suggested by many previous works [e.g., 6] for better performance. SIFT [4] is then adopted to describe each keypoint as a 128 dimensional vector. For each type of keypoints, we generate a visual vocabulary of 500 words using *k*-means. The visual words were found using features extracted from the web videos of TRECVID 2010 concept detection task. As processing all MED video frames will be computationally very expensive, we sample one frame from every two seconds.

Spatial-temporal interest points

While SIFT describes 2D local structures in images, spatial-temporal interest points (STIP) capture space-time volumes where the image values have significant local variations in both space and time. We use Laptev's method [7] to compute locations and descriptors for STIPs in video. The detector is based on an extension of Harris operator to space-time as described in [7]. Their code does not contain scale selection; instead interest points are detected at multiple spatial and temporal scales. HOG (Histograms of Oriented Gradients; 72 dimensions) and HOF (Histograms of Optical Flow; 72 dimensions) descriptors are computed for the 3D video patches in the neighborhood of the detected STIPs. We use concatenated HOGHOF feature (144 dimensions) as the final descriptor for each STIP.

MFCC audio feature

In addition to visual features like SIFT and STIP, audio is another important cue for detecting events in videos. We expect it to be complementary to the visual features. In audio processing, the mel-frequency cepstrum (MFC) is a representation of the short-term power spectrum of a sound. Mel-frequency cepstral coefficients (MFCC) are coefficients that collectively make up an MFC. They are derived from a type of cepstral representation of the audio clip (a nonlinear "spectrum-of-a-spectrum"). MFCC has been very popular in a large variety of audio related applications such as speech recognition. For TRECVID MED task, we compute MFCC feature over every 32ms time-window with 50% (16ms) overlap.

2.1 Bag-of-X Representation

Given a video clip, we extract three sets of features (SIFT, STIP, and MFCC). Due to the variations in length and content complexity, the sets of the same feature differ in cardinality across different video clips. This creates difficulties for learning methods (e.g., classifiers) which usually demand feature vectors of fixed dimension as input. To address this problem, we adopt vector quantization (VQ) technique to cluster each type of feature descriptors in feature space into a large number of clusters (i.e., "words") using the k-means clustering algorithm, and then encode each descriptor by the index of the cluster to which it belongs. The features extracted from each frame or audio time window are collapsed into a single bag. At the end, each video clip will be represented by a fixed dimensional histogram, one for each descriptor type. This forms the popular Bag-of-X framework which has been widely used for representing documents (bag-of-word), images (bag-of-visual-word), as well as audio sounds (bag-of-audio-word).

A fundamental difference of Bag-of-X representation for images or audio sounds is that the visual/audio words are the outcome of clustering algorithms (as opposed to the natural word entities in text documents). As a result, the performance of Bag-of-X image/sound representation is subject to several representation choices, such as vocabulary size (the number of clusters/words) and word weighting scheme in VQ.

Different from the SIFT feature for which 500-d vocabularies are chosen, we empirically generate a vocabulary of 4000 words for the STIP and MFCC feature. Smaller vocabularies are used for SIFT because we also applied spatial partitioning (a.k.a. spatial pyramid; Figure 2) of the video frames which prefers compact vocabularies [6, 8]. While for STIP, no spatial partitioning is used as we have found it unhelpful from our previous internal studies. For all the three features, we adopt a soft-weighting strategy for VQ which has been proved very useful for alleviating the quantization effect caused by clustering [9].

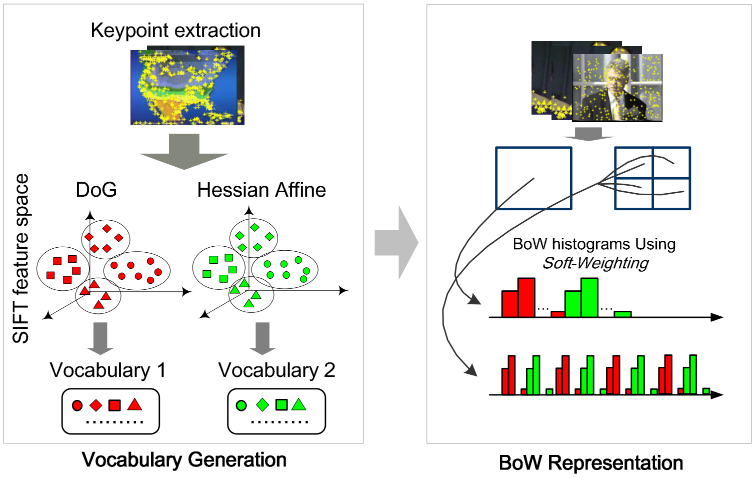


Figure 2: Image representation using static SIFT features with spatial partitioning (1x1 and 2x2). The two histograms are concatenated into a 5000-d feature vector as the final representation for each video frame. Given a video clip, we aggregate the 5000-d features from its sampled frames together as our clip-level feature representation. The clip-level features are then L-2 normalized before classification by SVM.

Baseline classifiers

With the three feature modalities (SIFT: 5000-d; STIP: 4000-d; MFCC: 4000-d), we train three baseline SVM classifiers using χ^2 kernel for each of the events over MED 2010 development set (1700+ web videos). The average fusion of probability predictions from the three SVM classifiers forms our baseline submission run6.

3. Classification with Temporal Matching

While accumulating all the SIFT/STIP/MFCC features from a video clip into a single feature vector seems a reasonable choice, it neglects the temporal information within the video clip. We therefore apply the earth mover's distance (EMD) [10] to measure video clip similarity, which was used in our previous work in [3]. We only adopt SIFT feature in this experiment, and thus each video clip is represented by a sequence of 5000-d bag-of-visual-word feature vectors (cardinality equals to the number of sampled frames). As shown in Figure 3, EMD computes the optimal flows between two sets of frames/features, producing the optimal match between both sets. Specifically, let a video clip be $P = \{(p_1, w_{p1}), \ldots , (p_m,w_{pm})\}$ of m frames, where p_i is the index of the i^{th} frame, and w_{pi} is the corresponding weight (uniformly set as $1/m$ in this experiment). To match P with another video clip $Q = \{(q_1, w_{q1}), \ldots , (q_n,w_{qn})\}$ of n frames, the EMD is computed as

$$\mathrm{EMD}(P, Q) = \Sigma_i \Sigma_j f_{ij}d_{ij} / \Sigma_i \Sigma_j f_{ij}, \quad (1)$$

where the ground distance d_{ij} between frames p_i and q_j is measured by the χ^2 distance of their corresponding bag-of-visual-word features. The flow f_{ij} representing the amount of weight transferred from frames p_i and q_j is optimized in EMD by minimizing the overall transportation workload $\Sigma_i\Sigma_j f_{ij}d_{ij}$, subject to the following constraints:

$$\begin{aligned} & f_{ij} \geq 0 \\ & \Sigma_j f_{ij} \leq w_{pi} \\ & \Sigma_i f_{ij} \leq w_{qj} \\ & \Sigma_i\Sigma_j f_{ij} = \min(\Sigma_i w_{pi}, \Sigma_j w_{qj}). \end{aligned} \quad (2)$$

The EMD is then used in a generalized form of Gaussian kernel for SVM classification:

$$K(P,Q)=\exp^{-\rho\mathrm{EMD}(P,Q)}, \quad (3)$$

where ρ is the kernel width parameter.

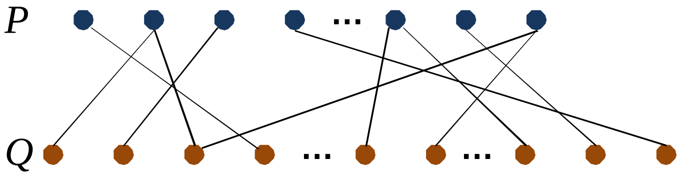


Figure 3: Toy example of EMD-based temporal matching between two sets of frames *P* and *Q*; lines indicate the presence of nonzero flows between corresponding frame pairs.

4. Diffusion with Generic Contextual Concept Detectors

Events are mostly defined by several (moving) objects such as "person", and generally occur under particular scene settings with certain audio sounds. For example, as shown in Figure 1, "batting a run in" contains people of various actions in the baseball field scene with typically some cheering or clapping sounds. Such event-scene-object-sound dependency provides rich contextual information for understanding the events. Most previous approaches, however, handled events, scenes, objects, and audio sounds separately without considering their relationship. Our intuition is that once the contextual cues can be computed, they can be utilized to make the event detection more robust. We therefore explore such contexts in MED 2010.

4.1 Building Contextual Concept Detectors

To build detectors (classifiers) for a large number of contextual concepts, we first need to collect enough training samples. To this end, we defined 21 contextual concepts as listed in Table 1, and designed an annotation tool to label the development videos (but not the test videos) for the 21 concepts. Each video is divided to multiple 10-sec clips, and the annotation is conducted at clip-level. Note that this annotation work is not just for this year's MED, but is a part of a long-lasting effort of building many contextual detectors for event detection.

With the labeled training data, we train SVM classifiers for detecting the concepts. For Human Action concepts, we use the bag-of-X representation of the STIP feature. Here we use a hierarchical k-means implementation to generate visual vocabularies of sizes 100, 200, 500, and 1000. Then a histogram intersection kernel multi-class SVM classifier is trained per action and vocabulary. Average fusion is used to combine classification outputs from different vocabulary sizes for each action. For the scene and audio concepts, we adopt similar classification framework based on the SIFT and MFCC features respectively. One difference in the audio concept classifiers is that we use a representation based on MFCC mean and covariance for each 10-sec clip, not the VQ histograms as in the event classifiers. From our evaluation, these context concept detectors worked fairly well. Figure 4 shows some top ranked frames (with high prediction scores) in the test video dataset.

Human Action Concepts	*Scene Concepts*	*Audio Concepts*
▪ Person walking	▪ Indoor kitchen	▪ Outdoor rural
▪ Person running	▪ Outdoor with grass/trees visible	▪ Outdoor urban
▪ Person squatting	▪ Baseball field	▪ Indoor quiet
▪ Person standing up	▪ Crowd (a group of 3+ people)	▪ Indoor noisy
▪ Person making/assembling stuffs with hands (hands visible)	▪ Cakes (close-up view)	▪ Original audio
		▪ Dubbed audio
▪ Person batting baseball		▪ Speech comprehensible
		▪ Music
		▪ Cheering
		▪ Clapping

Table 1: Contextual Concept Names.

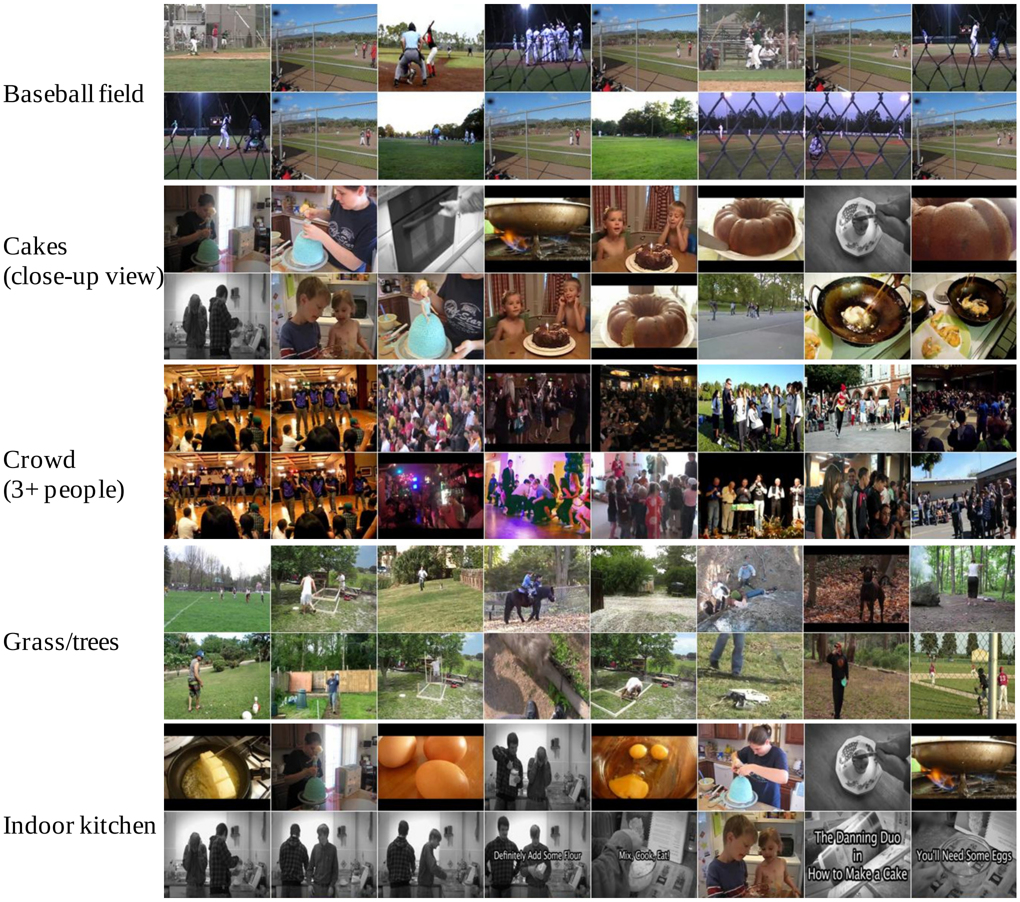


Figure 4: Top results of scene concept detection in MED 2010 test set. For each concept, frames of the top 16 detected video clips are shown, ordered from left to right and top to bottom.

4.2 Contextual Diffusion

To utilize these contextual detectors, we apply a contextual diffusion algorithm DASD (domain adaptive semantic diffusion) proposed in our prior work [11]. One underlying assumption of DASD is that detectors of frequently concurrent concepts/events should produce highly correlated scores. For example, the detection result of "baseball field" and "batting a run in" should be highly consistent as they frequently co-occur.

We therefore construct an undirected and weighted graph, namely semantic graph, to model the relationship between (and within) the events and the contextual concepts, where the relationship is estimated according to ground-truth labels of the events/concepts over the development dataset. In MED 2010, the graph contains 24 nodes (3 events and 21 contextual concepts). A part of the graph node relationship (only the event to concept relationship) is visualized in Figure 5. The graph is then applied to refine the detection scores using a function level diffusion process, where the aim is to recover the consistency of the detection scores w.r.t. the pair-wise relationship. More formally, the cost function of DASD is defined as:

$$E(\boldsymbol{g}, \mathbf{W}) = 1/2\ \Sigma_{ij}\ W_{ij}\ \|g(c_i) - g(c_j)\|^2 \qquad (4)$$

where $g(c_i)$ and $g(c_j)$ are the detection score vectors over a set of testing samples (video clips) for concepts/events c_i and c_j, and W_{ij} indicates the affinity (i.e., weight on the corresponding graph edge) between the two concepts/events.

Apparently, this cost function evaluates the smoothness of $\boldsymbol{g}$ over the semantic graph. Therefore, reducing the function value of E makes the detection results $\boldsymbol{g}$ more consistent with the concept affinities captured by $\mathbf{W}$. Specifically, we use gradient descent to reduce E by updating $\boldsymbol{g}$ iteratively. Interested readers are referred to [11] for more details of the DASD method. In the original DASD algorithm, it also involves a graph adaptation process which adapts $\mathbf{W}$ according to test data distribution. This process was not applied in MED 2010 as the development and test data are from similar domains.

	Indoor kitchen	Outdoor with grass/trees	Baseball field	Crowd (3+ people)	Cakes (close-up view)	Person walking	Person running	Person squatting	Person standing up	Making stuffs with hands	Person batting baseball	Indoor quiet	Indoor noisy	Outdoor rural	Outdoor urban	Original audio	Dubbed audio	Speech comprehensible	Music	Cheering	Clapping
Assembling a shelter	0.006	0.291	0.008	0.103	0.000	0.264	0.194	0.037	0.012	0.106	0.140	0.117	0.100	0.011	0.010	0.204	0.051	0.223	0.222	0.126	0.013
Battinga run in	0.000	0.093	0.694	0.078	0.000	0.116	0.047	0.001	0.000	0.043	0.002	0.020	0.000	0.282	0.338	0.072	0.308	0.117	0.106	0.007	0.638
Makinga cake	0.779	0.009	0.000	0.018	0.600	0.011	0.018	0.354	0.163	0.178	0.143	0.275	0.120	0.009	0.014	0.029	0.011	0.018	0.010	0.410	0.007

Figure 5: Estimated relationship (correlation) between the 3 events and the 21 contextual concepts according to ground-truth annotations. Note that the semantic graph not only models these event-concept relationships shown in this figure, but also the event-event and concept-concept correlations. The color-highlighted cells indicate the strong correlations discovered between events and the concepts.

5. Reranking with Event-Specific Object Detector

Besides the generic methods mentioned above, we are also interested in evaluating some ad-hoc ideas that are specific to individual events only. For event "batting a run in", videos usually contain certain human objects (e.g., batters) of familiar gestures and similar clothing. Assuming that videos of this event should have a high ratio of frames with batter visible, we trained a "batter" detector as an additional clue for detecting this specific event.

In order to train the detector, we manually labeled 800 positive frames in the development set by marking the bounding boxes of the whole body of batters. Negative samples (image patches) are randomly drawn from the other parts of the frames. Harr feature is extracted for each image patch. Other features such as HOG can also be applied but is not included in our current implementation. Training of the detector is based on the popular AdaBoost framework proposed in [12]. We found such specialized detectors to be reasonably effective in the small validation data set. Figure 6 shows a few examples of the detection results.

The batter detector is applied as a post-processing reranking step. For every test video, the ratio of frames that have positive detection of the batter object is first computed. If the ratio is larger than a threshold (its optimal value determined over a separate validation set), the event score of the video clip will be multiplied with the ratio and the video will be moved up to the top of the ranked list. Event detection scores of other videos are not modified.

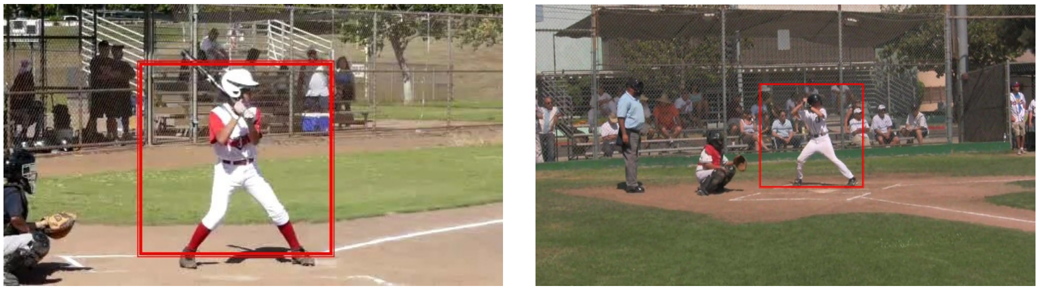


Figure 6: “batter” detection results.

6. Results and Analysis

Our MED system is designed to combine each of the core components introduced above. We submitted 6 runs to by incrementally adding in new components in order to study their effectiveness. Aside from the official submissions, we have also conducted evaluations over the dry-run validation set to analyze the contribution of each feature modality.

Figure 7 shows our six official submissions and all of the official TRECVID-2010 MED submissions. Performance is measured by mean Minimal Normalized Cost (MNC) over the three events. The MNC is computed based on the best (oracle) threshold of the detection scores, reflecting the best possible detection performance a system can reach. Detailed evaluation framework and description of the metrics can be found at http://www.nist.gov/itl/iad/mig/med10.cfm. From the figure we can see that a judicious approach using average fusion of the three feature modalities (Run6) already achieves very impressive results, with a mean MNC at 0.586. Incorporating the generic concept detectors (run4&5) using the graph diffusion algorithm provides moderate gains (run4 mean MNC 0.579). In addition, temporal matching with EMD kernel (run2&3) further improves the results (run2 mean MNC 0.565). The event-specific “batter” detector used in run1, however, didn’t prove useful from our current re-ranking tests. We conclude that event detection using multiple feature modalities is effective. While temporal matching with EMD kernel shows some noticeable gain, the contextual graph diffusion didn’t show significant improvements as we expected. This may be due to the fact that our event baseline has already used all three features combining different modalities (both visual and audio) In addition, our current implementation of concept detectors uses only single feature type (either visual or audio) and thus their performance may not be as strong as the multi-modal baseline detectors for events. Therefore, adding such relatively weak concept classifiers as context did not prove to be significantly beneficial in our current implementation.

Table 2 gives the performance of each feature component over the dry-run validation set. We see that all the three features perform fairly well, and the fusion of them significantly improves the results. Comparing the individual feature performance, STIP is slightly better than SIFT in terms of mean AP. This is consistent with the observations in recent works on action/event recognition [13]. For event “batting a run in”, SIFT outperforms STIP – which is probably because this event contains more consistent background scenes (“baseball field”), for which static SIFT features are very discriminative.

Another very interesting observation from our experiments is that standard audio feature MFCC, modeled in a bag-of-audio-word framework, demonstrates an impressive capability for event detection in unconstrained videos (though still lower than the visual counterparts). It is quite complementary to the state-of-the-art visual features used for event detection, as shown in the consistent accuracy improvement after fusing the audio-based detectors with the visual approaches. This shows the potential of jointly using both visual and audio features for multimedia event detection, which was only investigated in very few prior works [15].

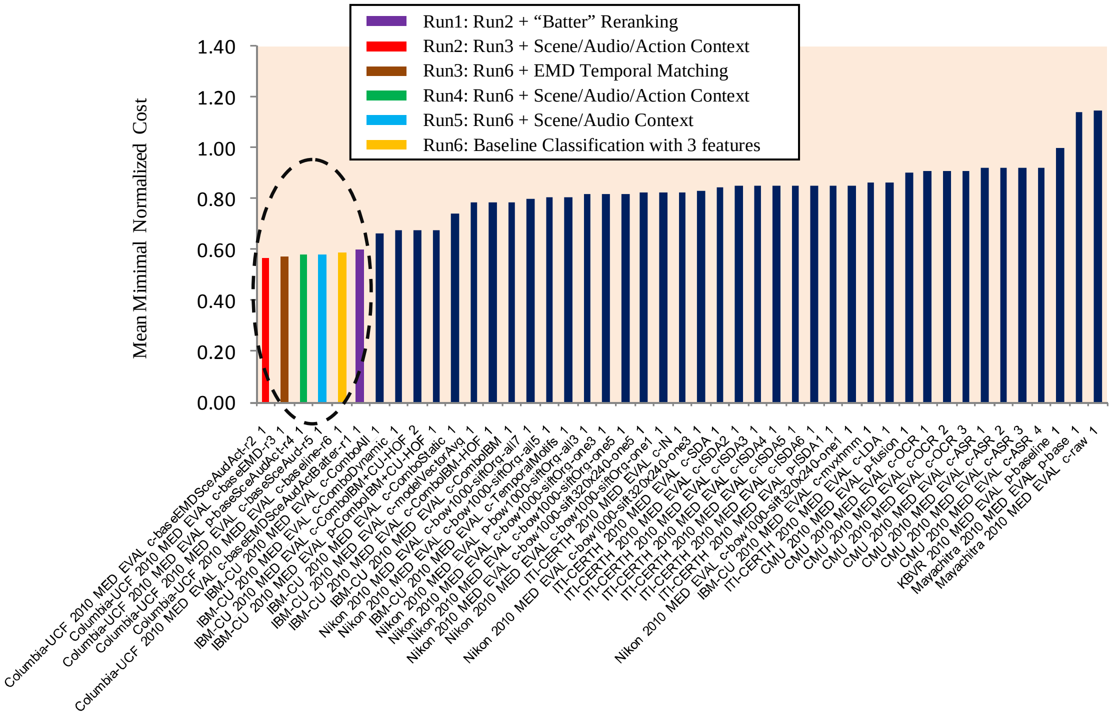


Figure 7: Performance of our submitted MED runs (circled) and all 45 official submissions. The vertical axis shows the performance measured by mean minimal normalized cost (mean MNC). This figure is best viewed in color.

	Assembling a shelter	Batting a run in	Making a cake	*Mean AP*
Visual STIP	0.468	0.719	0.476	0.554
Visual SIFT	0.353	0.787	0.396	0.512
Audio MFCC	0.249	0.692	0.270	0.404
STIP+SIFT	0.508	0.796	0.476	0.593
STIP+SIFT+MFCC	**0.533**	**0.873**	**0.493**	**0.633**

Table 2: Average precision (AP) performance of event detectors based on features of different modalities and their combinations. Note higher AP means better performance.

Figure 8 further displays the per-event performance of all the submissions. In addition to MNC, actual normalized cost (ANC) is also computed based on our provided threshold value. To determine the threshold, we simply treat the top 40 videos with the highest scores as positive and all the remaining ones as negative. In terms of both MNC and ANC, the runs we submitted demonstrate the best performance over all the submitted runs. In addition, from the results we also observed that different events favor different components or combination strategies. For example, the EMD-based temporal matching is helpful for "assembling a shelter" and "making a cake", but not for "batting a run in". Therefore another very interesting research direction is to investigate an adaptive method to automatically find out the best component and/or combinations for each event.

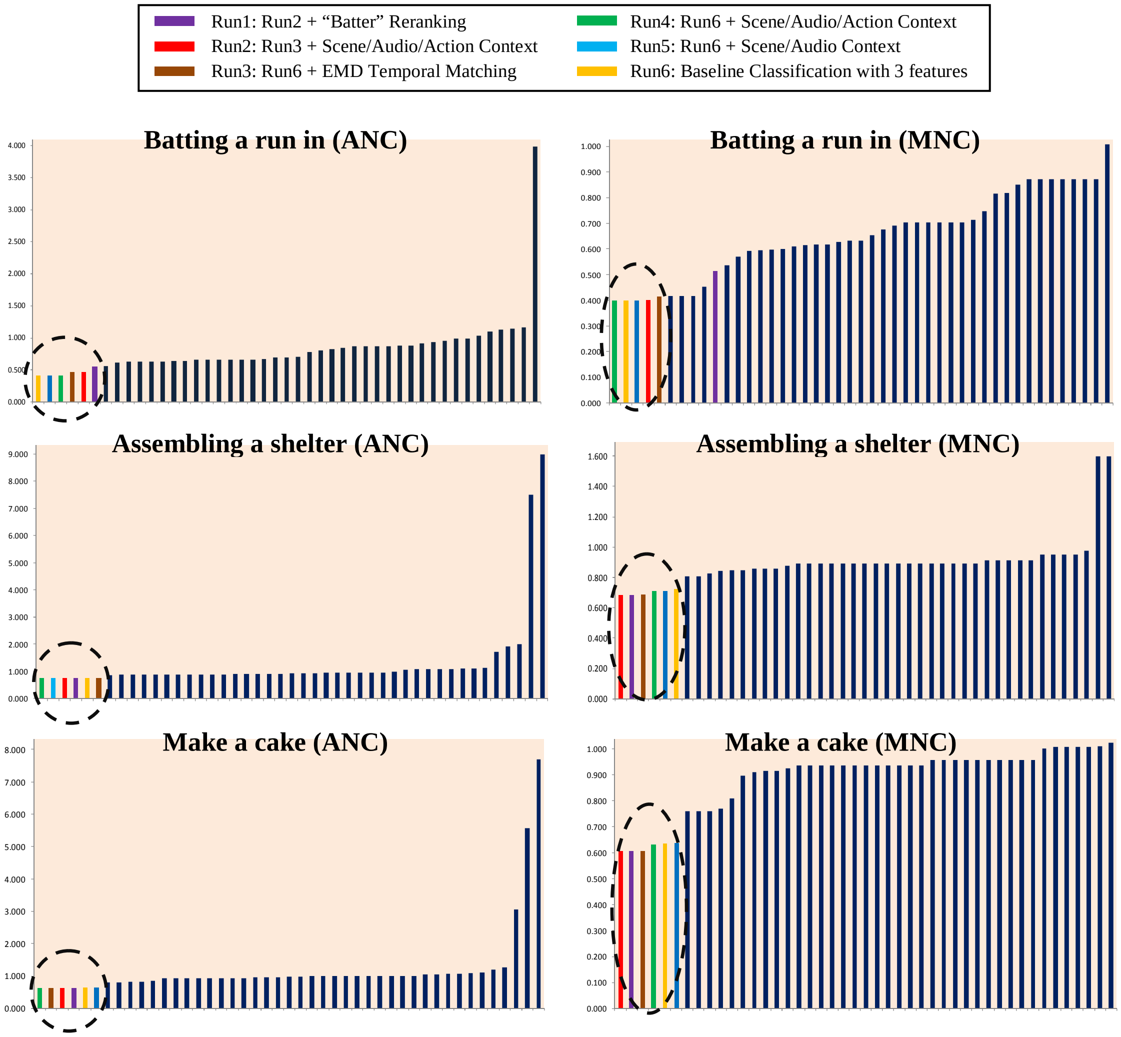


Figure 8: Per-event performance of our submitted MED runs (circled) and all 45 official submissions, measured by both minimal normalized cost (MNC) and actual normalized cost (ANC). We obtained top performance for all the three events. Note lower cost values mean better performance.

7. Acknowledgement

Many thanks to Yan Wang for developing the concept annotation tool and all the members of Columbia's Digital Video and Multimedia Lab for their help in the annotation effort. This work was supported in part by NGA Award # HM1582-09-1-0036, NSF Awards # CNS-07-16203 and CNS-07-51078.

8. References

[1] S.-F. Chang, D. Ellis, W. Jiang, K. Lee, A. Yanagawa, A.C. Loui, J. Luo, "Large-scale multimodal semantic concept detection for consumer video", ACM MIR, 2007.

[2] J. Liu, J. Luo, M. Shah, "Recognizing realistic actions from videos 'in the wild'", CVPR, 2009.

[3] D. Xu, S.-F. Chang, "Video event recognition using kernel methods with multi-level temporal alignment", IEEE Trans. on PAMI, vol. 30, no. 11, pp. 1985-1997, 2008.

[4] D. G. Lowe, "Distinctive image features from scale-invariant keypoints", IJCV, vol. 60, no. 2, pp. 91-110, 2004.

[5] K. Mikolajczyk, C. Schmid, "Scale and affine invariant interest point detectors", IJCV, vol. 60, no. 1, pp. 63-86, 2004.

[6] Y.-G. Jiang, J. Yang, C.-W. Ngo, A. G. Hauptmann, "Representations of keypoint-based semantic concept detection: a comprehensive study", IEEE Trans. on Multimedia, vol. 12, no. 1, pp. 42-53, 2010.

[7] I. Laptev, "On space-time interest points", IJCV, vol. 64, no. 2, pp. 107-123, 2005.

[8] S. Lazebnik, C. Schmid, J. Ponce, "Beyond bags of features: spatial pyramid matching for recognizing natural scene categories", CVPR, 2006.

[9] Y.-G. Jiang, C.-W. Ngo, J. Yang, "Towards optimal bag-of-features for object categorization and semantic video retrieval", ACM CIVR, 2007.

[10] Y. Rubner, C. Tomasi, L. J. Guibas, "A metric for distributions with applications to image databases", ICCV, 1998.

[11] Y.-G. Jiang, J. Wang, S.-F. Chang, C.-W. Ngo, "Domain adaptive semantic diffusion for large scale context-based video annotation", ICCV, 2009.

[12] P. Viola, M. Jones, "Rapid object detection using a boosted cascade of simple features", CVPR, 2001.

[13] M. Marszalek, I. Laptev, C. Schmid, "Actions in context", CVPR, 2009.

[14] J. Liu, Y. Yang, M. Shah, "Learning semantic visual vocabularies using diffusion distance", CVPR, 2009.

[15] W. Jiang, C. Cotton, S.-F. Chang, D. Ellis, A. C. Loui, "Short-term audio-visual atoms for generic video concept classification", ACM Multimedia, 2009.