Resource Constrained Multimedia Event Detection

Zhen-zhong Lan, Yi Yang, Nicolas Ballas, Shoou-I Yu, Alexander Haputmann

School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213, USA
lanzhzh,yiyang@cs.cmu.edu
nicolas.ballas@cea.fr
iyu,alex@cs.cmu.edu

Abstract. We present a study comparing the cost and efficiency trade-offs of multiple features for multimedia event detection. Low-level as well as semantic features are a critical part of contemporary multimedia and computer vision research. Arguably, combinations of multiple feature sets have been a major reason for recent progress in the field, not just as a low dimensional representations of multimedia data, but also as a means to semantically summarize images and videos. However, their efficacy for complex event recognition in unconstrained videos on standardized datasets has not been systematically studied. In this paper, we evaluate the accuracy and contribution of more than 10 multi-modality features, including semantic and low-level video representations, using two newly released NIST TRECVID Multimedia Event Detection (MED) open source datasets, i.e. MEDTEST and KINDREDTEST, which contain more than 1000 hours of videos. Contrasting multiple performance metrics, such as average precision, probability of missed detection and minimum normalized detection cost, we propose a framework to balance the trade-off between accuracy and computational cost. This study provides an empirical foundation for selecting feature sets that are capable of dealing with large-scale data with limited computational resources and are likely to produce superior multimedia event detection accuracy. This framework also applies to other resource limited multimedia analysis such as selecting/fusing multiple classifiers and different representations of each feature set.

Keywords: Multimedia Event Detection, Limited Resource, Feature Selection

1 Introduction

Multimedia data have proliferated in the past few years, ranging from ever-growing personal video collections to films and professional documentary archives. Numerous tools and applications have been invented to describe, organize, and manage video data. Previous research mainly focuses on recognizing scene, object and action, which are building blocks of events and defined as atomic concepts in this paper. However, these atomic concepts are too primitive to be used for

users to search videos from data collections. When searching through online video communities such as YouTube, Hulu etc., people tend to use event description such as "birthday party", "playing a board trick" or "mountain climbing" instead of simple scene, object or action words such as "indoor", "cake" or "walk". In this paper, we define an event as a combination of various actions, scenes and objects, which is more descriptive and meaningful.

The TRECVID MED evaluation [16], which is hosted by National Institute of Standards and Technology (NIST) and part of the TRECVID evaluation, is aimed at addressing above problems by assembling state-of-the-art technologies into a system that can quickly and accurately search a multimedia collection for user-defined events. Since 2010, NIST has collected one of the largest and most challenging labelled video datasets, which contains a total of 144049 video clips. These videos contain more than 30 events such as 'making a sandwich', 'parkour' and 'parade', which are illustrated in Fig. 1. Participants from various organizations have made significant progress on MED in terms of accuracy. However, most of the progress researchers have made comes from adding more and more features into their MED systems. While promising results can be achieved on such systems, they are too expensive to be deployed in real-world applications with large-scale data.

Fig. 1: Example Key-frame for Event in MED.

Another problem of TRECVID MED that has been discussed among participants and organizers for a long time is that NIST did not provide a validation set with labels, researchers from around the world publish MED related papers with their own splitting of TRECVID MED into testing and training sets. Because of these independent splits, comparing different research groups' results becomes very hard. To deal with these difficulties, NIST recently released two standard validation datasets, namely MEDTEST and KINDREDTEST. It is important to have some baseline results on these two datasets that can be compared by researchers from all over the world.

This paper attempts to address the above issues by thoroughly evaluating more than 10 multimedia features' performances and their contributions on the MEDTEST and KINDREDTEST datasets. Relying on this evaluation, we also propose a framework to select a subset of features to make a trade-off between accuracy and computational cost.

The remaining sections are organized as follows. We discuss related work in Section 2 and we elaborate our MED system including features, feature representations and evaluation metrics section 3. In Section 4, we discuss experimental results. Finally, we summarize our paper in Section 5.

2 Related Work

Compared with action recognition, recognizing a "complex event" is a new topic that has been introduced to take multimedia analysis to its next level of difficulty. Previous work [14] [17] [6] [13] on video event recognition can be divided in two main categories whether they rely on low-level features or high-level semantic concepts. Yang et al. [20] and Tamrakar et al. [17] proposed to evaluate the individual performance of different low-level visual features (SIFT, STIP, Trajectories...) as well as their combination. Meler et al. [14], Ebadollahi et al. [6] and Liu et al. [13] focus on testing high-level features performance on event recognition. In contrast to those works, this paper evaluates the performance of multi-modal features extracted from different streams associated with the multimedia data (image, audio, text), including both low-level and high-level features, leading to a more complete description. Moreover, most previous work only evaluates each feature's single performance. In contrast, we focus on each feature's contribution to the combined system. We show that the single feature performance, although important, do not necessarily reflect its contribution to the overall performance.

In terms of efficiency, most previous work focuses on improving one component of a classification or recognition system with faster algorithms. For example, Bay et al. [3] introduced SURF as an faster alternative of SIFT. Moosmann et al. [15] proposed to use random forest to replace Support Vector Machine (SVM). Jiang [7] conducted an interesting study to evaluate and combine a number of speed-up strategies to get a fast event recognition system. Different from previous work, we offer a resource constrained solution that can be customized by users who have different needs and resources.

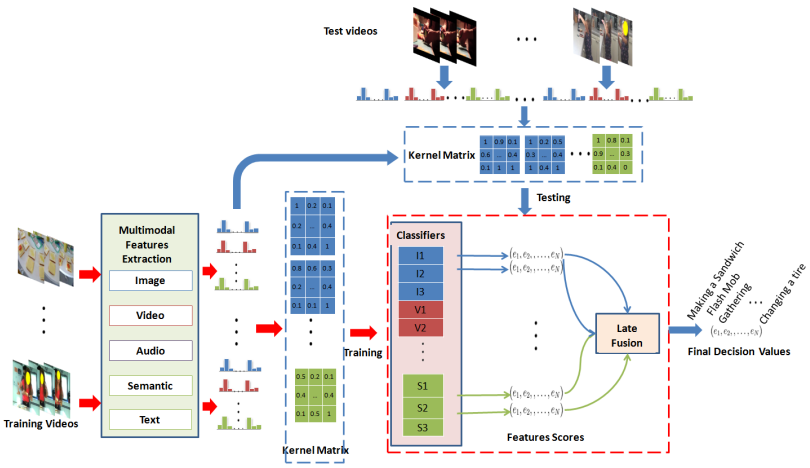


Fig. 2: MED system illustration.

3 MED System

Fig. 2 shows a simplified version of our MED system which is used for this paper. Given a set of training and testing videos, we first extract features in different modalities from the videos and then train a χ^2 SVM classifier for each feature. A simple average late fusion is used to combine the prediction results each feature. More complex fusion methods may lead to better performance, but this is beyond the scope of this paper: providing a baseline for NITS's two newly released MED datasets and illustrating a framework for designing resource constrained MED system.

3.1 Features

To build a good MED system, it is important to have features that capture various aspects of an event. In our MED system, we explore five different feature modalities which are computed from different sources. Image features capturing appearance information are computed from key-frames. Video features are extracted from videos directly and collect motion information. Audio features characterizes acoustic information. Text features and semantic features can borrow domain knowledge from other datasets such as Flickr and give semantically meaningful representations for events.

Image Features: We use three image features that are computed from the keyframes extracted as described in [10]. The three images feature are SIFT, Color SIFT (CSIFT) and Transformed Color Histogram (TCH) [18].

After detecting key points using harris-laplace key point detectors from key frames, we use three different feature descriptors to generate SIFT, CSIFT and TCH features, which hopefully are complementary. From the key points descriptors, a k-means algorithm generates a codebook which has 4096 words for each feature. Next, a soft-mapping strategy, in which we choose the ten nearest clusters and assign a rank weight ($\frac{1}{rank}$) for them, maps key points into the codebook. Spatial pyramid matching as described in [9] compensates for spatial information lost in the bag-of-words representation. We then aggregate the image representation into video representations by averaging all the image representations in one video and normalize the video representation using an L2 normalization.

Video Features: We have three visual video features, namely Dense Trajectory (Traj) [19], MoSIFT [4] and STIP [11], which are computed directly from videos. Traj is a feature that uses dense optical flow to track feature points up to 15 frames to get trajectories, which are described by Histogram of Oriented Gradient (HOG) [5], Histogram of Optical Flow (HOF) and Motion Boundary Histogram (MBH) [19]. By computing MBH along the dense trajectories, Traj has an efficient solution to compensate for camera motion. MoSIFT, as a three dimensional extension of SIFT features, uses a Difference of Gaussian (DoG) based detector and is represented by a descriptor combining SIFT and HOF. STIP uses 3D Harris corner detectors and its interest points are represented as the combination of HOG and HOF. After getting the key point descriptors,

the same bag-of words and spatial pyramid matching as with image features is adopted to cast key point representation into a video-level representation.

Audio Features: Audio features are another important resource to detect events in videos. To represent general audio information, we use the Mel-frequency cepstral coefficients (MFCCs) feature, which is very popular in speech recognition systems. We compute 20 dimensional MFCCs for every 10ms over a 32ms sliding window. Given the raw features, we compute a 4096 word codebook and aggregate all MFCC features from one video into a 4096 dimensional bag-of-words representation. In addition to MFCC, we also use have Automatic Speech Recognition (ASR) features as described in [2] to capture semantic information in audio.

Semantic Features: In our MED system, three semantic features are used. The first one called SIN346 is defined by the TRECVID Semantic Indexing (SIN) track. This feature has 346 dimensions representing the 346 concepts in SIN [2]. The second one is Object Bank feature (ObjBank) introduced by Li et.al. [12], in which we extended the original 176 objects to 1000 objects by using the Imagenet challenge 2012 dataset (ILSVRC2012) [8]. Another semantic feature that is also trained on the ILSVRC2012 dataset is the Deep Convolutional Neural Network feature (DCNN), in which we trained a Deep Convolutional Neural Network feature using the method introduced by Krizhevsky et al. [8] on a NVIDIA Tesla K20m GPU.

Text Features: Following Bao et al. [2], we also use Optical Character Recognition (OCR) features to represent the text feature. We use a commercial OCR system is used to recognize the text. As OCR rarely gets a complete word correct, we treat each trigram of characters as a token instead of each whole word as a token.

3.2 Evaluation Metrics

For performance comparison, three evaluation metrics are used: the first one is the Minimal Normalized Detection Cost(MinNDC) as indicated in Formula 1. It is an evaluation criteria for NIST to evaluate MED 2010 and MED 2011. Lower MinNDC indicates better performance.

$$NDC(S,E) = \frac{C_{MD} * P_{MD} * P_T + C_{FA} * P_{FA} * (1 - P_T)}{MINIMUM(C_{MD} * P_T, C_{MD} * (1 - P_T))} \quad (1)$$

where P_{MD} is the miss detection probability while P_{FA} is the false positive rate. $C_{MD} = 80$ is the cost for miss detection, $C_{FA} = 1$ is the cost for false alarm and $P_T = 0.001$ is a constant defining the priori rate of event instances.

Another metric that is related to MinNDC is $P_{MD}@TER = 12.5$, in which $TER = \frac{P_{MD}}{P_{FA}}$. Because MinNDC and $P_{MD}@TER = 12.5$ are only used for NIST evaluation and do not consider the ranking information or the detection result, we will also use mean average precision (MAP) as our evaluation criterion and use it to rank the performance of features because it is better at reflecting features' value due to its ranking sensitive characteristics.

4 Experiments

4.1 Data

We evaluate our system on two standard MED datasets, i.e., MEDTEST set and KINDREDTEST set, which both contain the same 20 events. The events and their ids are listed in Table 1. MEDTEST contains a total of 34051 video clips including 9094 training videos and 24957 testing videos. KINDREDTEST has the same training set but a testing set that only contains 12388 video clips.

Table 1: MED12 event ID and name.

E06: Birthday Party	E21: Attempting a bike trick
E07: Changing a vehicle tire	E22: Cleaning an appliance
E08: Flash mob gathering	E23: Dog show
E09: Getting a vehicle unstuck	E24: Giving directions to a location
E10: Grooming an animal	E25: Marriage proposal
E11: Making a sandwich	E26: Renovating a home
E12: Parade	E27: Rock climbing
E13: Parkour	E28: Town hall meeting
E14: Repairing an appliance	E29: Winning a race without a vehicle
E15: Working on a sewing project	E30: Working on a metal crafts project

4.2 Computing environment

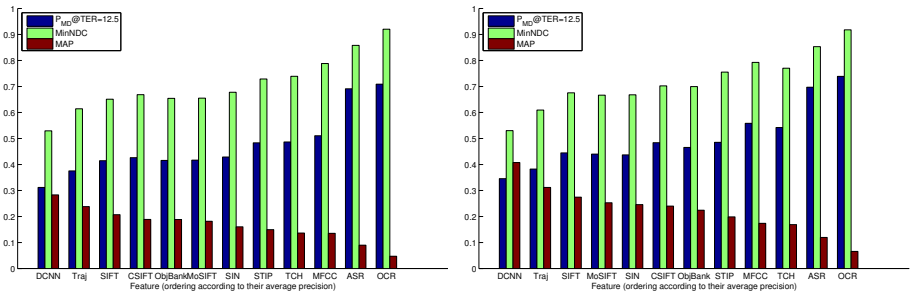
For extracting features, we use the PSC blacklight [1] machine, which is an SGI UV 1000cc-NUMA shared-memory system comprising 256 blades. Each blade holds 2 Intel Xeon X7560 (Nehalem) eight-core 2.27 GHz processors. Compared to feature extraction, the classification and fusion time is minimal, so we will only take the feature extraction time into consideration in this paper.

4.3 Single feature performance and contribution

Following the pipeline in Fig. 2, we study both single and combined features' performance using the three evaluation metrics described in Section 4.

Fig. 3 shows the single feature accuracies on our MEDTEST and KINDREDTEST sets. We order the accuracy according to their MAP. From Fig. 3, we can see that although the rank varies for different metrics and datasets, in general, they are consistent with each other. Specifically, the top two features that significantly outperform other features are DCNN and Traj; the two features that are worse than other features are ASR and OCR; others have very similar performances and the rank order changes are due to minor performance difference. It is interesting to see that DCNN, a high-level feature, can significantly outperform low-level features. Also, our OCR has higher recognition accuracy than ObjBank and SIN, yet its overall performance is the worst among all eleven

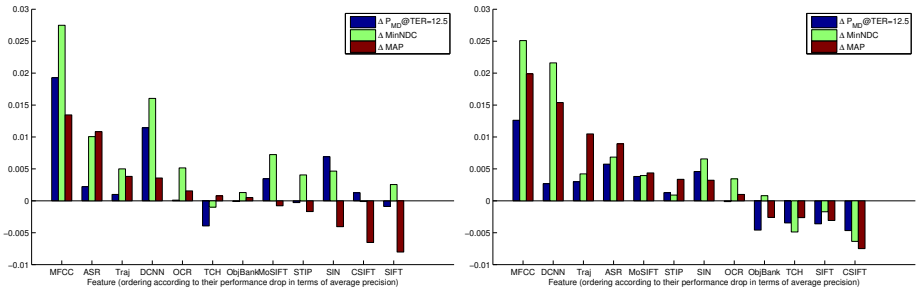
features. The reason is that these videos do not contain enough text to recognize. The big difference between visual and audio features shows that in unconstrained videos visual information is more distinctive than audio information.



MEDTEST KINDREDTEST

Fig. 3: Single feature accuracy for both datasets, ranked according to MAP. Lower score corresponds to better performance for MinNDC and $P_{MD}@TER = 12.5$, but higher is better for MAP.

To determine the contribution of each feature, we first calculate the performance by combining all features, then we remove one feature from the set and recalculate the performance. Fig. 4 shows the performance drop (leave-one-feature-out accuracy) from removing each feature. This drop shows the importance of each feature to the overall combined system. The ranking by performance drop is quite different than the ranking of single feature performance. These two rankings are statistically uncorrelated. For example, MFCC has a very poor ranking as a single feature accuracy but is the highest ranking for leave-one-feature-out performance. This indicates that MFCC is orthogonal to the other features. While SIFT and CSIFT, align with most other features, they reduce MAP because they reduce the overall weight per feature while not contributing additional information in the average late fusion method. More sophisticated fusion methods such as fusion by learning combination weights may be able to avoid this problem but will inevitably give smaller weights to those redundant features. Fig. 5 shows the Spearman rank coefficients for all of the features: it indicates MFCC and ASR very different from the other features. Although the Spearman rank coefficients also show OCR is very different from the other features, its close to random individual performance indicates its negligible role in the system. Fig. 4 demonstrates that single feature accuracy alone does not indicate suitability for inclusion in the combined feature set. However, as long as the leave-one-feature-out accuracy is not negative, inclusion will increase the overall score. Unfortunately, including all features with a positive value in Fig. 4 will lead to a computationally expensive system that is generally unsuitable for a real world applications.



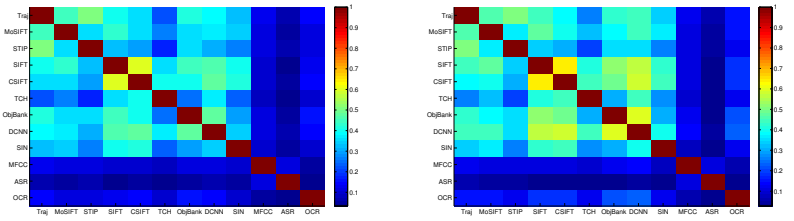
MEDTEST KINDREDTEST

Fig. 4: Leave-one-out Accuracy for MED, Ranked According to ΔMAP. In all three metrics, higher values means higher *performance drop* when we leave the feature out, hence a higher contribution of the feature to the combined system.

4.4 Performance versus cost trade-off

In order to determine the performance versus cost trade-off, we first determine each feature's computational cost as shown in Table 2, which also show the abbreviation of features for late usage in Table 3 through 6. For each feature, the time is the number of hours to process one hour of video. Let's assume our goal is to process one hour of video in one hour. We then determine, for a given number of CPUs, what the best possible performance is by a brute-force search across all features in Table 2. Fig. 6 shows the best possible performance for the given number of CPUs for all three metrics without using a GPU, excluding the DCNN feature. Tables 3 and 4 show the optimal feature sets for the given number of CPUs. Fig. 7 shows the best possible performance for the given number of CPUs for all three metrics using a GPU, including the DCNN feature. Tables 5 and 6 give the optimal feature sets for the given number of CPUs. As we can see from Tables 3 to 6, the MFCC feature appears in almost all configurations due to its low computational cost (Table 2) and relatively high contribution (Fig. 4). Although Traj has a high contribution, it does not show up in Tables 3 to 6 until we have a minimum of 16 CPUs due to its high computational cost. We can see from the tables that the optimal feature sets are very similar for the MEDTEST and the KINDREDTEST, which demonstrates that it is possible to select the optimal feature set from a smaller dataset like KINDREDEST and apply it to a larger dataset like MEDTEST. Likewise, these optimal feature sets are fairly similar across the three metrics. Further, we can also see from the figures that we get a diminishing return beyond 32 CPUs. In all cases, we can get more than 92 percent of the best performance by just using 32 cores.

Comparing Fig. 4 and Tables 5 and 6, we can see that in listing the importance of features, where importance in the table is measured by the ratio of the number of times the feature occurs to the number of possible occurrence given timing constraints, leave-one-feature-out performance is consistent with brute-force search results, hence very predictive in selecting the right feature set. For

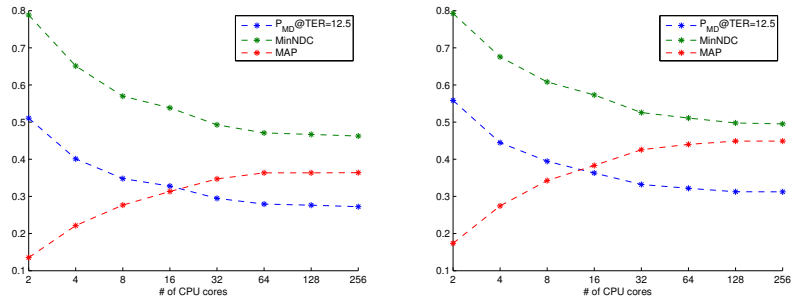


MEDTEST KINDREDTEST

Fig. 5: Spearman's rank correlation coefficient for features.

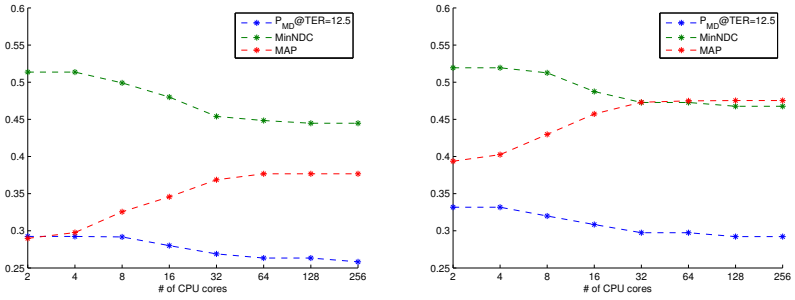
Table 2: Computational cost for features.

Features (Abbrev.)	core hours	features (Abbrev.)	core hours
Traj(Tr)	12.38	Objbank(Ob)	28.43
MoSIFT(Mo)	11.23	DCNN(DC)	0.15 GPU
STIP(ST)	10.33	SIN(SIN)	78.92
SIFT(SI)	3.57	MFCC(MF)	1.36
CSIFT(CS)	5.05	ASR(AS)	4.99
TCH(TC)	2.12	OCR(OCR)	1.34



MEDTEST KINDREDTEST

Fig. 6: Resource specific performance for MED (without DCNN).



MEDTEST KINDREDTEST

Fig. 7: Resource specific performance for MED (with DCNN on a GPU).

example, MFCC, DCNN, Traj and ASR, as the top 4 contributing features appear in almost all of the configurations as long as we have enough computational resources in terms of MAP. For other metrics, we have the same basic observation. The cost of computing the leave-one-feature-out accuracy is relatively inexpensive for late fusion, as all the components are already computed. In our system with 12 features, leave-one-feature-out accuracy computation is about 300 times faster than brute-force search.

Table 3: Resource specific feature sets for MEDTEST.

CPUs	Optimal Sets in Real-time Performance		
	MinNDC	$P_{MD}@TER = 12.5$	MAP
2	MF	MF	MF
4	TC MF	SI	TC MF
8	TC SI MF	TC SI MF	TC SI MF
16	Tr TC MF	CS SI AS MF	Tr TC MF
32	Mo TC CS SI AS MF	Mo TC CS SI OCR AS MF	Tr TC SI OCR AS MF
64	Ob ST Mo TC CS AS MF	Ob Mo Tr SI OCR AS MF	Ob Mo Tr TC OCR AS MF
128	Ob ST Mo TC CS OCR AS MF	Ob Mo Tr CS SI OCR AS MF	Ob Mo Tr TC OCR AS MF
256	SIN Ob Mo Tr CS SI OCR AS MF	SIN Ob Mo Tr SI OCR AS MF	SIN Ob Mo Tr TC OCR AS MF

Table 4: Resource specific feature sets for KINDREDTEST.

CPUs	Optimal Sets in Real-time Performance		
	MinNDC	$P_{MD}@TER = 12.5$	MAP
2	MF	MF	MF
4	SI	SI	SI
8	TC SI MF	TC SI MF	TC SI MF
16	Tr MF	Tr TC MF	Tr TC MF
32	ST Mo SI AS MF	Tr CS SI OCR AS MF	Tr TC SI AS MF
64	ST Mo Tr SI OCR AS MF	Ob Mo Tr SI OCR AS MF	ST Mo Tr TC SI OCR AS MF
128	SIN ST Mo Tr SI OCR AS MF	SIN ST Mo Tr SI OCR AS MF	SIN Mo Tr TC OCR AS MF
256	SIN Ob ST Mo Tr CS SI OCR AS MF	SIN Ob ST Mo Tr SI OCR AS MF	SIN Mo Tr TC OCR AS MF

Table 5: Resource specific feature sets for MEDTEST (with 1 additional GPU).

CPUs	Optimal Sets in Real-time Performance		
	MinNDC	$P_{MD}@TER = 12.5$	MAP
2	DC MF	DC MF	DC MF
4	DC MF	DC MF	DC TC MF
8	DC SI MF	DC SI MF	DC SI MF
16	DC Tr MF	DC TC SI OCR AS MF	DC Tr TC MF
32	DC Tr SI AS MF	DC Mo TC SI OCR AS MF	DC Tr CS OCR AS MF
64	DC Ob ST Mo TC SI AS MF	DC ST Mo Tr TC CS SI OCR AS MF	DC Ob Mo Tr TC OCR AS MF
128	SIN DC ST Mo TC SI OCR AS MF	SIN DC ST Mo SI OCR AS MF	DC Ob Mo Tr TC OCR AS MF
256	SIN DC Ob Mo Tr AS MF	SIN DC ST Mo SI OCR AS MF	DC Ob Mo Tr TC OCR AS MF

5 Conclusion

In this paper, we systematically evaluated the performance and contributions of more than 10 multi-modality features for complex event detection on uncon-

Table 6: Resource specific feature sets for KINDREDTEST(with 1 additional GPU).

CPUs	Optimal Sets in Real-time Performance		
	MinNDC	$P_{MD}@TER = 12.5$	MAP
2	DC MF	DC MF	DC MF
4	DC MF	DC MF	DC TC MF
8	DC AS MF	DC AS MF	DC SI MF
16	DC Tr MF	DC Tr MF	DC Tr MF
32	DC Tr AS MF	DC Mo Tr OCR AS MF	DC Tr SI AS MF
64	DC Tr AS MF	DC Mo Tr OCR AS MF	DC Mo Tr SI AS MF
128	SIN DC ST Mo Tr SI OCR AS MF	SIN DC Mo Tr OCR AS MF	SIN DC Mo Tr TC OCR AS MF
256	SIN DC ST Mo Tr SI OCR AS MF	SIN DC Mo Tr OCR AS MF	SIN DC Mo Tr TC OCR AS MF

strained videos over two newly released TRECVID MED datasets: MEDTEST and KINDREDTEST. These results can serve as a baseline for the community.

Based on the evaluation and computational cost of feature extraction, we propose a resource constrained video analysis framework that can meet different users' needs. More specifically, we select feature sets that have optimal real-time performance under various resource constraints by measuring leave-one-feature-out performance and brute-force search performance.

A particularly important insight from above experiments is that leave-one-feature-out feature performance is very predictive in selecting the right feature set.

We also found that in both datasets and across the three different metrics:

- DCNN and Trajectory features are very useful features in unconstrained video analysis. Especially DCNN, given its semantic and high accuracy characteristics, is a feature that is worth paying a lot of attention to.
- Even a less accurate feature such as MFCC, if it is cheap and complementary to other features, can be very useful.
- By selecting the right features, we can save a large amount of computational cost with a minimum accuracy drop. For example, in our experiments, by reducing the computational cost of 87 percent we still achieve 92 percent of optimal performance.

5.1 Acknowledgements

This work was partially supported by Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center contract number D11PC20068. 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. This material is based in part upon work supported by the National Science Foundation under Grant No. IIS-1251187.

References

1. http://www.psc.edu/index.php/computing-resources/blacklight.
2. Lei Bao, Shoou-I Yu, Zhen-zhong Lan, Arnold Overwijk, Qin Jin, Brian Langner, Michael Garbus, Susanne Burger, Florian Metze, and Alexander Hauptmann. Informedia@ trecvid 2011. TRECVID2011, 2011.
3. Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. Surf: Speeded up robust features. In ECCV, pages 404–417. Springer, 2006.
4. Ming-yu Chen and Alexander Hauptmann. Mosift: Recognizing human actions in surveillance videos. CMU-CS-09-161, 2009.
5. Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In CVPR, volume 1, pages 886–893. IEEE, 2005.
6. Shahram Ebadollahi, Shih-fu Chang Xie, Lexing, and Smith John R. Visual event detection using multi-dimensional concept semantics. ICME, pages 881–884, 2006.
7. Yu-Gang Jiang. Super: Towards real-time event recognition in internet videos. In ICMR, page 7. ACM, 2012.
8. Alex Krizhevsky, Ilya Sutskever, and Geoff Hinton. Imagenet classification with deep convolutional neural networks. In NIPS, pages 1106–1114, 2012.
9. Zhen-zhong Lan, Lei Bao, Shoou-I Yu, Wei Liu, and Alexander G Hauptmann. Double fusion for multimedia event detection. In MMM, pages 173–185. Springer, 2012.
10. Zhen-zhong Lan, Lei Bao, Shoou-I Yu, Wei Liu, and Alexander G Hauptmann. Multimedia classification and event detection using double fusion. Multimedia Tools and Applications, pages 1–15, 2013.
11. Ivan Laptev. On space-time interest points. IJCV, 64(2-3):107–123, 2005.
12. Li-Jia Li, Hao Su, Li Fei-Fei, and Eric P Xing. Object bank: A high-level image representation for scene classification & semantic feature sparsification. In NIPS, pages 1378–1386, 2010.
13. Jingen Liu, Qian Yu, Omar Javed, Saad Ali, Amir Tamrakar, Ajay Divakaran, Hui Cheng, and Harpreet S Sawhney. Video event recognition using concept attributes. In WACV, pages 339–346, 2013.
14. Michele Merler, Student Member, Bert Huang, Lexing Xie, and Gang Hua. Semantic Model vectors for complex video event recognition. IEEE Trans. on Multimedia, 14(1):88–101, 2012.
15. Frank Moosmann, Eric Nowak, and Frederic Jurie. Randomized clustering forests for image classification. PAMI, 30(9):1632–1646, 2008.
16. Paul Over, George Awad, Martial Michel, Jonathan Fiscus, Greg Sanders, Barbara Shaw, Wessel Kraaij, Alan F. Smeaton, and Georges Quenot. Trecvid 2012 – an overview of the goals, tasks, data, evaluation mechanisms and metrics. In TRECVID. NIST, USA, 2012.
17. Amir Tamrakar, Saad Ali, Qian Yu, Jingen Liu, Omar Javed, Ajay Divakaran, Hui Cheng, Harpreet Sawhney, and S R I International Sarnoff. Evaluation of low-level leatures and their combinations for complex event detection in open source videos. CVPR, pages 3681–3688, 2012.
18. Koen EA Van De Sande, Theo Gevers, and Cees GM Snoek. Evaluating color descriptors for object and scene recognition. PAMI, 32(9):1582–1596, 2010.
19. Heng Wang, Alexander Klaser, Cordelia Schmid, and Cheng-Lin Liu. Action recognition by dense trajectories. In CVPR, pages 3169–3176. IEEE, 2011.
20. Jun Yang, Yu-Gang Jiang, Alexander G Hauptmann, and Chong-Wah Ngo. Evaluating bag-of-visual-words representations in scene classification. In Workshop on ICMR, pages 197–206. ACM, 2007.