IBM Research TRECVID-2004 Video Retrieval System

Arnon Amir[3], Janne O Argillander[2], Marco Berg[3], Shih-Fu Chang[4], Martin Franz[2], Winston Hsu[4], Giridharan Iyengar[2], John R Kender[1], Lyndon Kennedy[4], Ching-Yung Lin[1], Milind Naphade[1], Apostol (Paul) Natsev[1], John R. Smith[1], Jelena Tesic[1], Gang Wu[1], Rong Yan[1], Donqing Zhang[1]

[1] IBM T. J. Watson Research Center	[2] IBM T. J. Watson Research Center	[3] IBM Almaden Research Center	[4] Columbia University E.E. Department
19 Skyline Drive	Yorktown Heights,	650 Harry Rd	New York, NY 10027
Hawthorne, NY 10532	NY 10598	San Jose, CA 95120	

Abstract

In this paper we describe our participation in the NIST TRECVID-2004 evaluation. We participated in four tasks of the benchmark including shot boundary detection, high-level feature detection, story segmentation, and search. We describe the different runs we submitted for each track and provide a preliminary analysis of our performance.

1. Introduction

Content-based retrieval of video presents significant challenges in terms of development of effective techniques for analysis, indexing and searching of video databases. TRECVID is greatly facilitating the advancement of technologies for content-based retrieval of video by providing a standard dataset and evaluation forum for evaluating emerging and novel techniques and systems. The IBM team participated in TRECVID for the fourth time since its inception in 2001. The goal of our participation in 2004 was to participate in all four of the TRECVID tasks – shot boundary detection, high-level feature detection, story segmentation, and search (manual and interactive) – and to explore large variation of techniques for each task. As a result, we developed a wide range approaches and systems, and we submitted the maximum number of runs for each task.

2. Shot Boundary Detection

The IBM team participated in the shot boundary determination (SBD) task for TRECVID 04 and submitted ten runs. The IBM CueVideo system was used, which was explored in prior years at TRECVID. More details of the SBD system and analysis of the results will be provided in the final paper.

3. High-Level Feature Detection

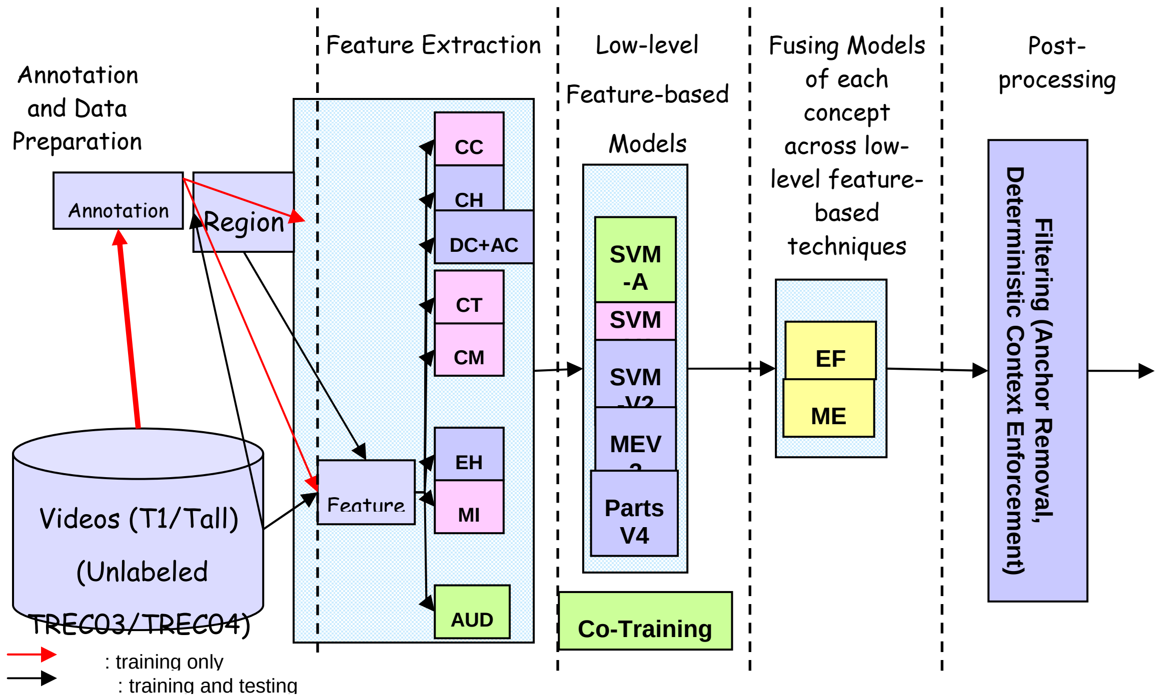
The IBM team participated in the high-level feature detection task for TRECVID 04 and submitted ten runs.

3.1. The IBM TRECVID 2004 Concept Detection System

The TRECVID 2004 Concept Detection Task included 10 concepts (or high level features), most of which are rare in terms of frequency of occurrence in the training set. The IBM system this year was therefore geared to this challenge of rare concept detection.

The System consists of the feature extraction modules, for regional and global visual features as well as text-based features from the Automatic Speech Recognition and/or Closed Caption Text made available to the participants by NIST. We experimented with visual features extracted from the compressed stream directly as well as those extracted from the decompressed keyframes. This was followed by the feature-based modeling modules. We tried mainly two approaches, one based on support vector machine classification and the other based on maximum entropy based classification. The SVM modeling used various compressed-domain based and decompression-based visual and text features. The maximum entropy approach used a similar set of visual features. Visual features included color Correlograms, histograms, edge histograms, color moments, wavelet texture, co-occurrence texture, moment invariants etc. A validation set based scheme was used to tune classifier parameters. We then fused the outputs of different models based on combinations of features and classifiers using two techniques: ensemble fusion and maximum entropy. We then applied deterministic contextual filtering to remove anchor shots, vary shot relevance based on shot length and position within the broadcast etc. Unlike the IBM TRECVID 2003 Concept Detection System Pipeline, the Context Enforcement Module was not enforced in the 2004 system As the concepts in the benchmark this year were rare and we found that the common annotation set was not annotated with enough level of detail, a lot of context that could have been learnt and used for enforcement was missing from the training set annotations. Based on this pipeline we had various combinations of processing modules to create 10 runs.

The IBM TRECVID 2004 Concept Detection Pipeline is shown in the Figure below



To approach the problem of rare occurrence of most benchmark concepts in the training set and to utilize the multiple modalities in a systematic fashion, we experimented with a novel approach to leverage unlabeled data sets in conjunction with labeled data sets and to combine the multiple modalities. We refer to this approach as CFEL or cross feature ensemble learning. All ten of the runs we submitted combined at least 1 visual model output with one output from the text-based model. All runs that combine model outputs from all 4 models for visual features (SVM-V1, SVM-V2, MEV, and Parts) are referred to as "Mall". All runs that used the training samples from the feature development corpus of TRECVID 2003 are referred to as "Tall". All runs that leveraged an unlabeled data set along with the available labeled data sets have the prefix "CM" in their run name. 8 of the runs did not have any filtering stage applied. The table below lists the name of the IBM run and its description.

- BOM: Best combination of single A and V

- Mall_T1_EF: All models, Ensemble Fusion
- Mall_T1_MEMF: All models, ME Fusion
- Mall_Tall_EF: All models, all sets, Ensemble Fusion
- CM2all_T1_EF: All models, Co-training, Ensemble Fusion
- CM2all_T1_MEMF: All models, Co-training, ME Fusion
- (TREC03 set as unlabeled set)
- CM2all_Tall_EF: All models, Co-training, All sets, EF (TREC03 set as unlabeled set)
- CM4all_Tall_EF: All models, Co-training, All sets, EF (TREC04 set as unlabeled set)
- Filter1: Mall_T1_EF filtered (w/anchor, depth filtered for 2 concepts)
- Filter2: CM2all_Tall_EF filtered (w/anchor, depth filtered for 2 concepts)

3.2. Concept Detection Results

The IBM System has once again topped in mean average precision across all the 82 submitted runs. IBM runs have resulted in topmost average precision performance for 4 of the 10 concepts, topmost precision at 100, 1000 and 2000 for 5 of the 10 concepts. Figure 1 compares IBM performance with the best performance across all runs for the 10 concepts. Figure 2 compares the IBM runs' precision at a depth of 100.

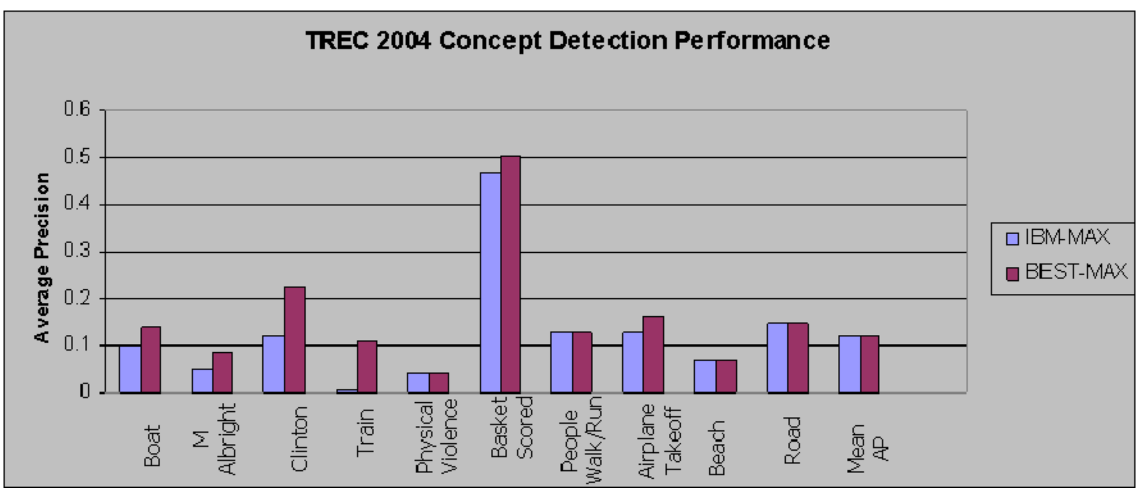


Figure 1. TRECVID 2004 Concept Detection Performance (average precision).

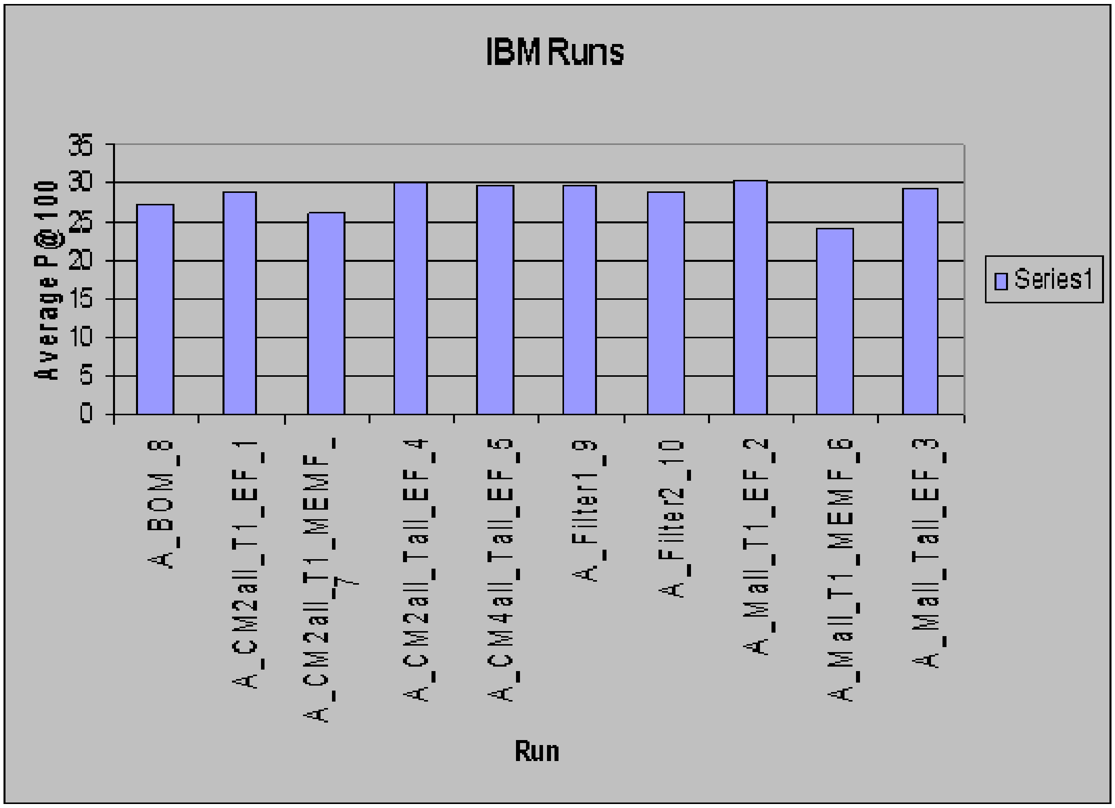


Figure 2. TRECVID 2004 Concept Detection Performance (average precision).

3.3. *Concept Detection Lessons*

The following lessons were learnt from across all ten IBM runs submitted:

1. Multimodal fusion improves over any single modality significantly.
2. Cross-Feature Ensemble Learning helps improve precision towards the top
3. Except 1 run all other runs improve over BOM
4. Maximum Entropy failed as a fusion strategy and resulted in worse performance then Ensemble Fusion
5. Filtering improves one or two concepts due to anchor removal but not substantially.
6. Use of TREC03/TREC04 as unlabeled set in co-training gives almost similar results.
7. SVM classifiers worked better than others for rare class classification

4. Story Segmentation

Recent research in video analysis has shown a promising direction, in which mid-level features (e.g., people, anchor, indoor) are abstracted from low-level features (e.g., color, texture, motion, etc.) and used for discriminative classification of semantic labels. However, in most systems, such mid-level features are selected manually. In this story boundary segmentation experiment, we propose an information-theoretic framework, visual cue cluster construction (VC^3), to automatically discover adequate mid-level features that are relevant to story boundary detection. The problem is posed as mutual information maximization, through which optimal cue clusters are discovered to preserve the highest information about the semantic labels. We extend the Information Bottleneck framework to high-dimensional continuous features and further propose a projection method to map each video into probabilistic memberships over all the cue clusters. The biggest advantage of the proposed approach is to remove the dependence on the manual process in choosing the midlevel features and the huge labor cost involved in annotating the training corpora for training the detector of each mid-level feature. The proposed VC^3 framework is general and effective, leading to exciting potential in solving other problems of semantic video analysis. When tested in news video story segmentation, the proposed approach achieves promising performance gain over representations derived from conventional clustering techniques and even our prior manually-selected mid-level features.

In this experiment, we focus on the automatic approach to construct salient mid-level visual cue clusters. To exploit the support from other modalities, we also fuse other features, such as speech prosody features [19] and ASR-based segmentation scores [13], which have been shown highly relevant to story boundary detection in the broadcast news videos.

4.1. Approach

News videos from different channels usually have different production rules or dynamics. We choose to construct a model that adapts to each different channel. According to our research in [23], the fusion capability of the discriminative model, such as SVM, is more effective than other generative or ensemble models. We train a SVM classifier to classify a candidate point as a story boundary or non-boundary. The features fed to the SVM classifier include the membership probabilities of the induced VC^3 clusters, ASR based segmentation scores [13]., and speech prosody features [19].

As our prior work [19], we take the union of shot boundaries and audio pauses as candidate points but remove duplications within 2.5-second fuzzy window. Our study showed these two sets of points account for most of the story boundaries in news.

4.1.1. Visual Cue Cluster Construction (VC^3)

In the research of video retrieval and analysis, a new interesting direction is to introduce "mid-level" features that can help bridge the gap between low-level features and semantic concepts. Examples of such mid-level features include location (indoor), people (male), production (anchor), etc., and some promising performance due to such mid-level representations have been shown in recent work of news segmentation and retrieval [25][26]. It is conjectured that mid-level features are able to abstract the cues from the raw features, typically with much higher dimensions, and provide improved power in discriminating video content of different semantic classes. However, selection of the mid-level features is typically manually done relying on expert knowledge of the application domain. Once the mid-level features are chosen, additional extensive manual efforts are needed to annotate training data for learning the detector of each mid-level feature.

Our goal is to automate the selection process of the mid-level features given defined semantic class labels. Given a collection of data, each consisting of low-level features and associated semantic labels, we want to discover the mid-level features automatically. There is still a need for labeling the semantic label of each data sample, but the large cost associated with annotating the training corpus for each manually chosen mid-level feature is no longer necessary. In addition, dimensionality of the mid-level features will be much lower than that of the low-level features.

Discovery of compact representations of low-level features can be achieved by conventional clustering methods, such as K-means and its variants. However, conventional methods aim at clusters that have high similarities in the low-level feature space but often do not have strong correlation with the semantic labels. Some clustering techniques, such as LVQ, take into account the available class labels to influence the construction of the clusters and the associated cluster centers. However, the objective of preserving the maximum information about the semantic class labels was not optimized.

Recently, a promising theoretic framework, called Information Bottleneck (IB), has been developed and applied to show significant performance gain in text categorization [27]. The idea is to use the information-theoretic optimization methodology to discover "cue word clusters" which can be used to represent each document at a mid level, from which each document can be classified to distinct categories.

The cue clusters are the optimal mid-level clusters that preserve the most of the mutual information between the clusters and the class labels.

In the visual features of this experiment, we propose new algorithms to extend the IB framework to the visual domain, specifically video. Starting with the raw features X such as color, texture, and motion of each shot, our goal is to discover the cue clusters C^* that have the highest mutual information or the least information loss about the final class labels Y, such as video story boundary or semantic concepts. Note that X and Y are given in advance. The optimization criteria given some constrain R under IB principle can be form as the following:

$$C^* = \arg\min_{C|R} \left\{ I(X;Y) - I(C;Y) \right\}.$$

Our work addresses several unique challenges. First, the raw visual features are continuous (unlike the word counts in the text domain) and of high dimensions. We propose a method to approximate the joint probability of features and labels using kernel density estimation (KDE) [28]. Second, we propose an efficient sequential method to construct the optimal clusters and a merging method to determine the adequate number of clusters. Finally, we develop a rigorous analytic framework to project new video data to the visual cue clusters by taking into account the cluster prior probabilities and feature likelihood toward the specific cue clusters. The probabilities of such projections over the cue clusters are then used for the final discriminative classification of the semantic labels or fused with other modalities such as text or speech prosodies. More details regarding the VC^3 framework are addressed in [24].

4.1.2. Support Vector Machines (SVM)

SVM has been shown to be a powerful technique for discriminative learning [29]. It focuses on structural risk minimization by maximizing the decision margin. We applied SVM using the Radial Basis Function (RBF) as the kernel, $K(x_i, x_j) = \exp(-\gamma \| x_i - x_j \|), \gamma > 0$.

In the training process, it is crucial to find the right parameters C (tradeoff on non-separable samples) and γ in RBF. We apply five fold cross validation with a grid search by varying (C, γ) on the training set to find the best parameters achieving the highest accuracy.

4.2. Boundary Detection Results

We compare the VC^3 -based results with our previous work in Figure 3, where performance breakdowns are listed for different types of video stories (as defined in Figure 1 of [23]**Error! Reference source not found.**). The left-most bar is the percentage of data in each story type (among the 759 stories of 22 annotated CNN video programs). The second bar is the recall rate achieved by our prior work [19], where specific classifiers of anchors, commercials, and prosody-related features are fused through the Maximum Entropy approach, denoted as ME. Note that the ME approach uses both the audio and visual features. The VC^3 feature is shown in the third bar from the left, and includes visual features only. It is interesting that even with an automatic method to discover the salient visual features, the VC^3 approach can match our previous method (ME) that uses a specific detector like anchor, and uses features from multiple modalities. Meanwhile, the new VC3 adaptive approach also improves detection performance of some story types by being able to discover unique visual features such as the second anchor in story type (b), specific station animations in type (f), weather stories in type (h), and transition from the end of story to commercials in type (i).

As the case in our prior work [19], we further add other features such as prosody to the visual-only approach to achieve more performance improvement. The prosody features have been shown to be effective for detecting syntactically meaningful phrases and topic boundaries [19][19][30]. By adding similar feature set, we can see the improvement for several story types, especially when topic changes are correlated with by appearance of reporters or anchors such as type (a). Those challenging types such as type (c), multiple stories within a single anchor shot, or type (e), continuous short briefings without studio or anchor shots, benefit the most from adding the prosody features. The improvements can be clearly seen in the right-most bar, denoted by A+V, in Figure 3.

4.3. Discussions for Story Boundary Detection

We have proposed an information-theoretic VC^3 framework, based on the Information Bottleneck principle, to associate continuous high-dimensional visual features with discrete target class labels. We utilize VC^3 to provide new representation for discriminative classification, feature selection, and prune "non-informative" feature clusters. The proposed techniques are general and effective. Most importantly, the framework avoids the manual procedures to select features and greatly reduces the amount of annotation in the training data. More details of the theoretical framework of VC^3 and some related experiments can be found in [24].

Some extensions of VC^3 to induce audio cue clusters, support multi-modal news tracking and search are underway. Other theoretic properties such as automatic bandwidth selection for KDE and performance optimization are also being studied.

The fusion from other modalities, such as speech prosody features and ASR-based segmentation scores are significant and have been confirmed again in this experiment.

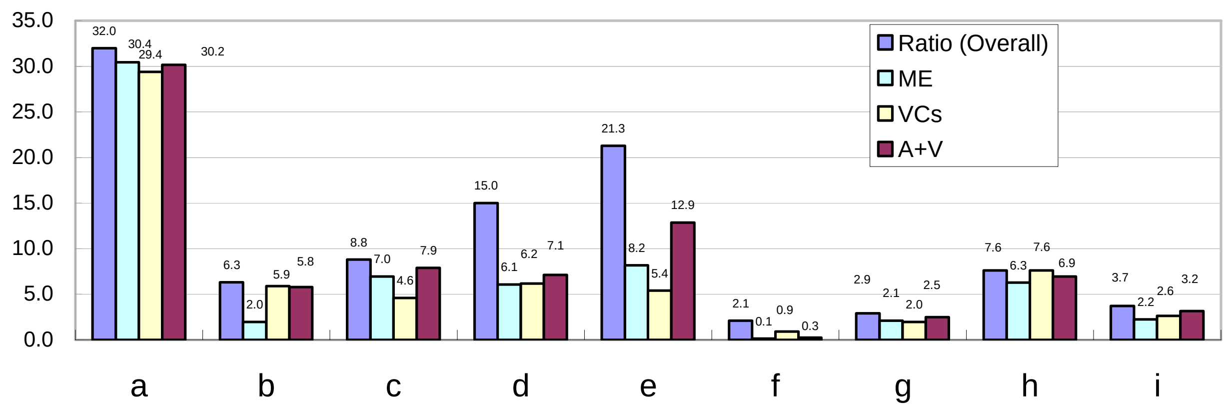


Figure 3. Story boundary detection performance comparisons of our previous approach (ME) [19][19], VCs only, and VCs combined with speech prosody features. The performance results over different story types (as defined in Figure 1 of [23]) are listed separately. The ME approach uses both audio and visual modalities; VCs uses visual cue clusters only. The 'ratio' group indicate is the percentage of storied in each type. All the numbers shown are the recall values, with the precision of each experiment fixed at 0.71.

5. Search

The IBM team participated in the search task for TRECVID 04 and submitted ten runs based on automatic, manual and interactive search. We participated in the Search task, submitting 3 interactive, 6 manual, and 1 fully automatic runs. We describe some of these runs below. More information and detailed analysis will be provided in the final paper.

5.1. Automatic Search

Our fully automatic search run was the combination of an automatic speech-based run and an automatic visual run. The speech run was based on the LIMSI ASR transcript [2] and the available Closed Caption text, using the alignment provided by CMU. Simple pre-processing—such as removal of stop words and the phrase "Find (more) shots of"—was performed to the query topic text in order to extract query keywords for each topic. The automatic visual run was based on the Multi-Example Content Based Retrieval (MECBR) approach used in the IBM automatic search run from last year [1]. Overall, this fully automatic run had a Mean Average Precision score of 0.057 which is higher than many of the manual runs and is virtually the same as the average MAP score (0.06) across all 67 manual and automatic runs. Changes from last year included a new set of visual features, a new visual query example selection method, and a late feature fusion method for combining query results for multiple feature hypotheses.

5.1.1. Feature selection and fusion

The approach adopted for feature selection was to optimize globally the feature type and granularity within each feature modality (e.g., color and texture), to perform early feature fusion in each independent modality, and late fusion across modalities. The motivation was that even though the relative importance of one feature modality vs. another (e.g., color vs. texture) may change from one topic to the next, the relative performance of the specific features within a given feature modality (e.g., color correlogram vs. color histogram) should be the same across all topics, and can therefore be optimized globally for all query topics. We therefore performed off-line experiments using the TRECVID 2003 query topics to select the best color feature type, granularity, and color feature combination, as well as the same parameters for the best texture feature. Based on the experiments, we selected the normalized combination (i.e., concatenation) of a global 166-dimensional HSV color correlogram and a 3x3 grid-based 81-dimensional Lab color moments feature as the best color feature. Similarly, we selected the normalized combination of a global 96-dimensional co-occurrence texture feature and a 3x3 grid-based 27-dimensional Tamura texture feature as the best overall texture feature. The third feature modality we used was that of 46-dimensional semantic model vectors built from the detection confidence scores with respect to 46 frequently occurring concepts.

5.1.2. Example selection and fusion

Visual query examples were selected using the following method. Each of the example video clips was processed to extract all I-frames in the clip and up to 3 of them were selected as representative clip keyframes. The boundary frames for each clip (e.g., the first 5 I-frames and the last 5 I-frames) were removed from consideration in order to avoid selecting shot transition frames. The remaining I-frames were sampled uniformly to select up to 3 visual query examples for the given video clip. All of the image examples, as well as the selected keyframes from each video clip, were used as independent content-based retrieval queries in each of the 3 feature spaces (color, texture, and semantic model vectors). The query results across all examples were normalized to 0 mean and unit standard deviation, and were fused using MAX score aggregation, essentially mimicking an OR logic for fusion across query examples (i.e., a good match to any of the examples was considered a good match overall).

5.1.3. Modality fusion

Given the retrieval scores for each of the four independent modalities (text, color, texture, and semantic model vectors), the range normalized scores were combined using a weighted average score aggregation, where the modality weights were proportional to the Mean Average Precision scores of the corresponding modality as measured on the TRECVID 2003 search topics. The specific weights used for text, color, texture, and semantic model vectors were 11, 4, 3, and 2, respectively.

5.2. Manual Search

5.2.1. Manual multi-modal TJW run

This run was generated using a query-specific combination of content-based retrieval (CBR), model-based retrieval (MBR), and simple text search (i.e., keyword spotting) based on the LIMSI ASR transcript. Each query was manually formulated as a Boolean or a weighted average combination of queries based on visual examples, semantic models, and/or speech keywords. The system used to generate this run supports a variety of visual features extracted at global, spatial layout-based and regular grid-based granularities. The set of features includes 166-d HSV color histogram, 166-d HSV color correlogram, 6-d

Lab color moments, 108-d Lab color wavelets, 96-d co-occurrence texture, 12-d wavelet texture, 3-d Tamura texture, 64-d edge histograms, and 6-d Dudani shape moment invariants. The system also supports retrieval based on higher-level semantic features as well as simple keyword matching in the speech transcript. This run had a fairly low MAP score of 0.048 which is primarily due to the simplicity of the speech retrieval model used.

5.2.2. Manual multi-modal ARC run

This run was generated from a multi-modal video retrieval system developed at the IBM Almaden Research Center. It relies primarily on speech-based retrieval and re-ranking based on the visual features described above. This run had the highest MAP score (0.109) among the IBM manual runs.

5.2.3. Manual visual-only run

We submitted one visual-only manual run which was generated similarly to the fully automatic visual run described above but with manually selected visual query examples. This run used the same visual features (color, texture, and model vectors) and the same example fusion method but a slightly different score normalization and aggregation method for fusion across the three visual feature modalities. In particular, the results were rank-normalized and fused with MAXAVG score aggregation in order to avoid scaling issues and bias towards any of the feature spaces. The MAXAVG score aggregation method essentially takes the maximum confidence score as the final aggregated score across the three features, and breaks score ties using the average of the three individual scores. The MAX score aggregation is a more liberal fusion method than averaging, and mimics an OR logic for fusion across modalities (i.e., a match in any of the modalities is considered a match overall), while the tie-breaking was necessary due to the large number of overall score ties resulting from the rank-based normalization of the individual scores. This run was basically an automatic run but with manually selected examples and it was submitted to evaluate the effect of manual example selection and rank-based feature fusion as compared to automatic example selection with weighted average feature fusion. Analysis of the results is still under way, however, since this was the only purely visual run and is not directly comparable to any of the other submitted runs (internal evaluation and comparison of other visual-only runs is in progress). This run was also used to generate late fusion-based variations of two other multi-modal manual runs, as described below.

5.2.4. Multi-modal fusion runs

We submitted two manual runs which were the result of late fusion between the visual-only run described above and the two primary manual runs (i.e., the multi-modal TJW run and multi-modal ARC run). The fusion method was identical in both cases, namely that of weighted average score aggregation with query-specific weights. For each query, the weights for the two runs were selected manually (based on the query topic description only) from among the following weight combinations (modulo symmetries): {0.1, 0.9}, {0.3, 0.7}, {0.5, 0.5}. Prior to score aggregation, the scores were normalized using linear range normalization. Unfortunately, in both cases, the fusion with the visual-only run actually hurt the overall performance, although there were improvements for several individual queries (6 topics improved in the fusion with the multi-modal ARC run and 8 topics improved in the fusion with the multi-modal TJW run). The two multi-modal fusion runs had MAP scores of 0.045 and 0.080, compared to the 0.048 and 0.011 scores for the two primary manual runs. The analysis of the results is ongoing but the poor fusion performance is likely due to the subjective way of setting the fusion weights, which were not derived or validated either empirically or visually. A more careful weight selection based on query type classification with pre-computed optimal query type weights, for example, could perhaps preserve the gains in the more visual queries without deteriorating the performance for the other queries.

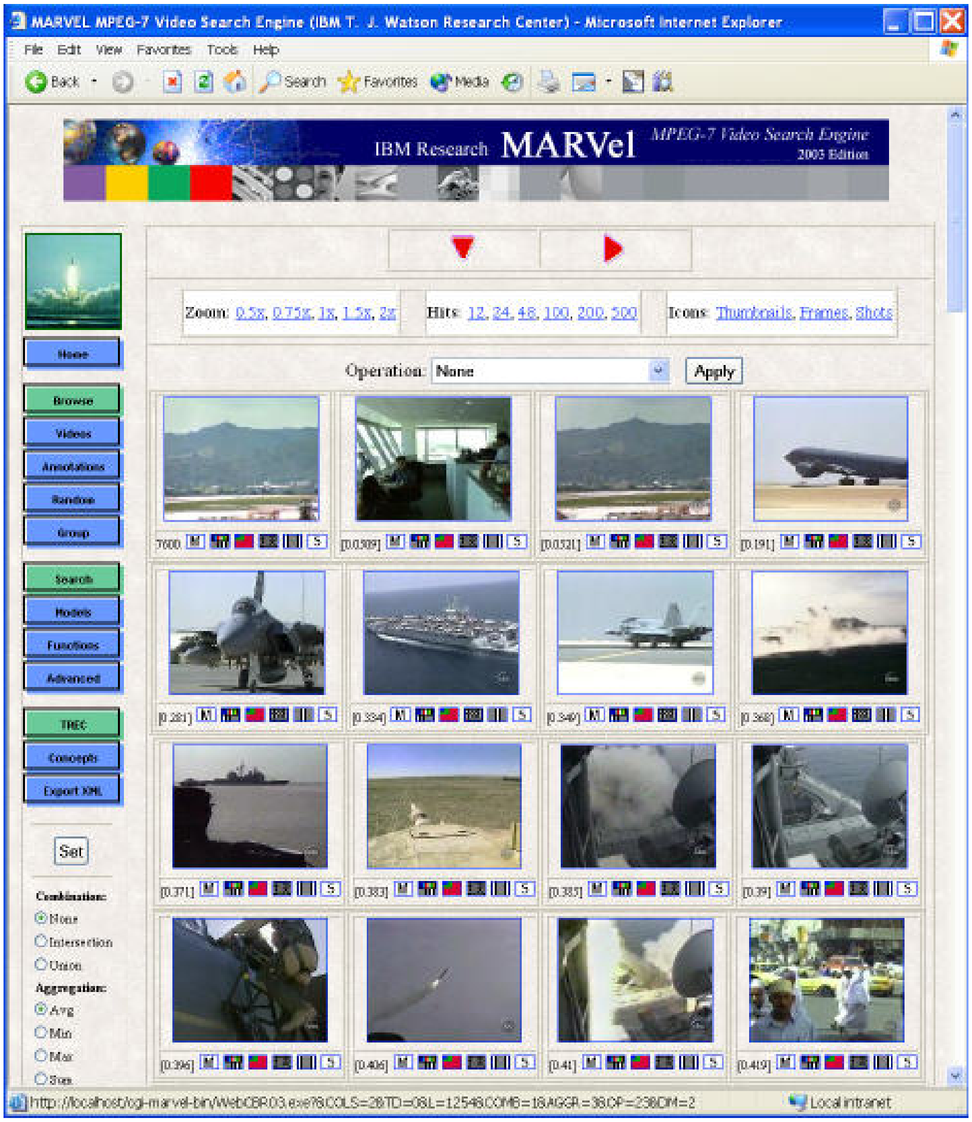
5.3. Interactive Search

5.3.1. Interactive multi-modal Almaden runs

An interactive system based on IBM CueVideo was explored for interactive search. An interactive multimedia retrieval system often provides both searching and browsing capabilities, using various multimodal indexes, feedback methods and manual shots selection and elevation. This is in contrast to a manual search task, which allows searching and query refinement (on other data sets) but not any browsing, nor manual shots selection. It is imperative that a user has to split the interactive task time between query refinement and manual shots browsing. An important question is how much do we gain from browsing and from manual shot elevation, compared to making a better query. Search and browse are complementary to each other in several different ways. In a typical Interactive session, a search is first performed, followed by a quick browsing over the results list. Next, the user may either refine the query and search again, or expand the browsing in neighborhoods developed around correct matches. A single search may find the proximity of many different correct matches, whereas browsing allows to pin-point the correct shots in each such proximity. Hence a search may be considered as a global operation over the entire database while browsing operates in small neighborhoods. While a search may requires a well formulated query, browsing needs only an initial reference point in a browsing space. This space could be video-ID and time like in a traditional storyboard, color and texture histograms in common content-based "show me more like this" browsing, or any other proximity criteria on which more correct matches are expected to be found. However, a good search query may capture many correct matches in a single, scalable operation over a large video corpus, while browsing does not scale well to large collections, especially when many correct matches exist. Our preliminary experiment with TRECVID-03 data and topics suggests that browsing and shots elevation plays a very important role in Interactive Search, in particular for rare topics. More details of the Almaden interactive search system and analysis of the results will be provided in the final paper.

5.3.2. Interactive multi-modal TJW run

An interactive system based on IBM MARVEL MPEG-7 video search was also explored for interactive search. The MARVEL search engine provides tools for content-based, model-based and speech term-based querying. The MARVEL search engine was used for one of the interactive search runs. The system allows the user to fuse together multiple searches within each query. The interactive search run typically used this capability for answering the query topics. For example, the user would typically examine the query topic and example content and then issue multiple searches based on the example content, models and speech terms. The IBM MARVEL MPEG-7 video search engine demo can be accessed at http://mp7.watson.ibm.com/.



6. Summary

In this paper we described our participation in the NIST TRECVID-2004 evaluation and discussed our approaches and results in four tasks of the benchmark including shot boundary detection, high-level feature detection, story segmentation, and search.

7. References

[1] A. Amir, S. Srinivasan, and D. Ponceleon. “Efficient video browsing using multiple synchronized views.” In Azriel Rosenfeld, David Doermann, and Daniel DeMenthon, editors, *Video Mining*. Kluwer Academic Publishers, Boston, USA, 2003.

[2] A. Natsev, J. R. Smith, “Active Selection for Multi-Example Querying by Content,” *Proc. IEEE Intl. Conf. on Multimedia and Expo (ICME)*, Baltimore, MD, July, 2003.

[3] A. Natsev, M. Naphade, J. R. Smith, “Exploring Semantic Dependencies for Scalable Concept Detection,” *Proc. IEEE Intl. Conf. on Image Processing (ICIP)*, Barcelona, ES, Sept., 2003.

[4] B. L. Tseng, C.-Y. Lin, M. Naphade, A. Natsev and J. R. Smith, “Normalized Classifier Fusion for Semantic Visual Concept Detection,” *IEEE Intl. Conf. on Image Processing*, Barcelona, Sep. 2003.

[5] C.-Y. Lin, B. L. Tseng and J. R. Smith, “Video Collaborative Annotation Forum: Establishing Ground-Truth Labels on Large Multimedia Datasets”, *Proc. of NIST TREC Video 2003*.
[6] C.-Y. Lin, B. L. Tseng, and J. R. Smith. “VideoAnnEx: IBM MPEG-7 Annotation Tool for Multimedia Indexing and Concept Learning.” *Proc. IEEE Intl. Conf. on Multimedia and Expo (ICME)*, Baltimore, MD, July, 2003.
[7] C.-Y. Lin, B. L. Tseng, M. Naphade, A. Natsev, J. R. Smith, “VideoAL: A end-to-end MPEG-7 video automatic labeling system,” *Proc. IEEE Intl. Conf. on Image Processing (ICIP)*, Barcelona, ES, Sept., 2003.
[8] E. D. Barraclough77. On-line searching in information retrieval, *Journal of Documentation*, 33:220-238, 1977.
[9] H. Lee and A. F. Smeaton. Designing the user interface for the Fischlar digital video library. *Journal of Digital information*, Special Issue on Interactivity in Digital Libraries, Vol.2:4, Article No. 103, 2002-05-21, 2002.
[10] H. Nock, W. Adams, G. Iyengar, C-Y Lin, M. R. Naphade, C. Neti, B. Tseng, J. R. Smith, "User-trainable Video Annotation using Multimodal Cues", *Proc. SIGIR 2003*
[11] J. R. Smith, A. Jaimes, C.-Y. Lin, M. Naphade, A. Natsev, B. L. Tseng, “Interactive Search Fusion Methods for Video Database Retrieval,” *Proc. IEEE Intl. Conf. on Image Processing (ICIP)*, Barcelona, ES, Sept., 2003.
[12] J. R. Smith, M. Naphade, A. Natsev, “Multimedia Semantic Indexing using Model Vectors,” *Proc. IEEE Intl. Conf. on Multimedia and Expo (ICME)*, Baltimore, MD, July, 2003.
[13] M. Franz, J. S. McCarley, S. Roukos, T. Ward, and W.-J. Zhu, “Segmentation and detection at IBM: Hybrid statistical models and two-tiered clustering broadcast news domain," in *Proc. of TDT-3 Workshop*, 2000.
[14] M. Naphade and J. R. Smith, "A Hybrid Framework for Detecting the Semantics of Concepts and Context", 2nd *Intl. Conf. on Image and Video Retrieval*, pp. 196-205, Urbana, IL, June 2003
[15] M. Naphade and J. R. Smith, "Learning Visual Models of Semantic Concepts", *IEEE International Conference on Image Processing*, Barcelona 2003
[16] M. Naphade, I. Kozintsev and T. Huang, “A Factor Graph Framework for Semantic Video Indexing”, *IEEE Transactions on Circuits and Systems for Video Technology* Jan 2002.
[17] M. Naphade, J. R. Smith, “Learning Regional Semantic Concepts from Incomplete Annotations,” *Proc. IEEE Intl. Conf. on Image Processing (ICIP)*, Barcelona, ES, Sept., 2003.
[18] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*, chapter 10: User Interfaces and Visualization, pages 257{323. Addison Wesley, Reading, USA, 1999.
[19] W. Hsu, S.-F. Chang, C.-W. Huang, L. Kennedy, C.-Y. Lin, and G. Iyengar, “Discovery and fusion of salient multi-modal features towards news story segmentation,” in *IS&T/SPIE Electronic Imaging*, San Jose, CA, 2004.
[20] W.H. Adams, A. Amir, C. Dorai, S. Ghosal, G. Iyengar, A. Jaimes, C. Lang, C.-Y.Lin, A. Natsev, C. Neti, H. J. Nock, H. Permuter, R. Singh, J. R. Smith, S. Srinivasan, B. L. Tseng, AT Varadaraju, D. Zhang,"IBM Research TREC-2002 Video Retrieval System,” *NIST Text Retrieval Conference (TREC-2002)*, Nov., 2002.
[21] IBM Team, “IBM Research TRECVID-2003 Video Retrieval System”, *NIST Text Retrieval Conference (TREC-2003), Gaithersburg*, MD, Nov. 2003
[22] J. L. Gauvain, L. Lamel, and G. Adda, “The LIMSI Broadcast News Transcription System”, *Speech Communication*, 37(1-2):89-108, 2002. ftp://tlp.limsi.fr/public/spcH4_limsi.ps.Z
[23] W. Hsu, S.-F. Chang, "Generative, discriminative, and ensemble learning on multimodal perceptual fusion toward news video story segmentation," ICME, Taipei Taiwan, 2004.
[24] Winston H. Hsu, S.-Fu Chang, “Visual cue cluster construction via information bottleneck principle and kernel density estimation,” International Conference on Content-Based Image and Video Retrieval, July, 2005, Singapore.
[25] Chaisorn, L., Chua, T.S., Koh, C.K., Zhao, Y., Xu, H., Feng, H., Tian, Q., "A two-level multi-modal approach for story segmentation of large news video corpus," TRECVID Workshop, 2003.

[26] Amir, A., Berg, M., Chang, S.F., Iyengar, G., Lin, C.Y., Natsev, A., Neti, C., Nock, H., Naphade, M., Hsu, W., Smith, J.R., Tseng, B., Wu, Y., Zhang, D., "Ibm research trecvid 2003 video retrieval system," TRECVID Workshop, 2003.
[27] N. Slonim, N. Tishby, “Agglomerative information bottleneck,” NIPS, 1999.
[28] D. W. Scott, “Multivariate Density Estimation: Theory, Practice, and Visualization,” Wiley-Interscience, 1992.
[29] V.N. Vapnik, "Statistical Learning Theory," Wiley, New York, 1998.
[30] E. Shriberg, A. Stolcke, D. Hakkani-Tur, and G. Tur, “Prosody-based automatic segmentation of speech into sentences and topics,” *Speech Communication*, vol. 32, pp. 127-154, 2000.