Columbia University TRECVID-2005
Video Search and High-Level Feature Extraction

Shih-Fu Chang, Winston Hsu, Lyndon Kennedy, Lexing Xie,
Akira Yanagawa, Eric Zavesky, Dong-Qing Zhang

Digital Video and Multimedia Lab, Columbia University
http://www.ee.columbia.edu/dvmm

Oct. 28th 2005

Descriptions of Submitted Runs

High-Level feature extraction

- **A_DCON1_1**: Choose the best-performing classifier from the following runs for each concept.
- **A_DCON2_2**: linear weighted fusion of 4 SVM classifiers using color/texture, parts-base classifier, and tf-idf text classifier.
- **A_DCON3_3**: same as above, except a new input-adaptive fusion method was used.
- **A_DCON4_4**: average fusion of 4 SVM classifiers using color/texture, parts-based classifier, and naïve Bayesian text classifier.
- **A_DCON5_5**: same as above, except the naïve Bayesian text classifier was not fused.
- **A_DCON6_6**: Choose the best performing uni-model visual classifier among the 4 SVM's and parts-based classifier for each concept. No classifier fusion was done.
- **A_DCON7_7**: a single SVM classifier using color and texture trained on the whole training set (90 videos) with 20% negative samples.

Search

- **I_C_2_ColumbiaI1_1**: Interactive system, using text search (against ASR/MT), content-based image retrieval, story-based browsing, text search against visual concepts, and near-duplicate detection.
- **I_C_2_ColumbiaI2_2**: Interactive system using text search (against ASR/MT transcript), content-based image retrieval, story-based browsing, text search against visual concepts, near duplicate detection, and Cue-X re-ranking.
- **I_C_2_ColumbiaI3_3**: Interactive run where an annotator spends the entire duration annotating the output from an automatic run. The automatic run uses query-class-dependent weights on text retrieval, content-base image retrieval, and text search against visual concepts .
- **M_C_2_ColumbiaM1_4**: Manual run using ASR/MT text retrieval, content-based image retrieval, and text search against visual concepts.
- **F_C_2_ColumbiaA1_5**: Multimodal automatic run. Uses query-class-dependent weights on text retrieval, content-based image retrieval, text search against visual concepts, and Cue-X reranking
- **M_A_1_ColumbiaM2_6**: Required manual baseline run, using only text searches against the ASR/MT transcript.
- **F_A_1_ColumbiaA1_7**: Required automatic baseline run, using only text searches against the ASR/MT transcript.

We participated in two TRECVID tasks in 2005 – "High-Level Feature Extraction" and all three types of "Search". Summaries of our approaches, comparative performance analysis of individual components, and insights from such analysis are presented below.

Task: High-Level Feature Extraction

In TRECVID 2005, we specifically explored the potential of parts-based statistical approaches in detecting generic concepts. Parts-based object representation and its related statistical detection models [1] have gained great attention in the computer vision community in recent years. This is evidenced by promising results reported in conferences like CVPR, ICCV, and NIPS. We analyzed their performance and compared them with some of the state of the art known from the TRECVID feature detection results in the previous years. We adopted a general approach and applied the same technique to all of the 10 concepts. One of our main objectives was to understand what types of high-level features would benefit most from such new representation and detection paradigm.

For the baseline method, we adopted the SVM-based method over two simple visual features – color moments over 5x5 fixed grid partitions and Gabor texture from the whole frame. We limited the features to the above two, fixed the grid partitions, and did not include other classification models (e.g., co-training). Such baseline technique, although simple, has been shown competitive in past TRECVID experiments and literature [2]. In the TRECVID 2005 results, it actually achieved a quite satisfactory performance with an MAP of 0.266 (within the 25% margin from the best MAP).

The parts-based paradigm nicely complements the above baseline approach. It represents each image as an attributed relational graph (ARG), with each node corresponding to a salient part (e.g., corner, high-entropy patch) that is automatically extracted from the image at different scales. It captures the local attributes associated with each part as well as the spatial relationships among the parts. Given a collection of training images and their extracted parts, a Random ARG model is then learned with machine learning techniques to maximize the data generation likelihood. Intuitively, the parameters of the Random ARG model, once learned, are able to capture the underlying statistical properties of the part attributes and the inter-part topological and attributive relationships. Such properties tend to correspond to the uncertainty caused by photometric, geometric, or imaging condition variations of objects in the real world. To enhance the classification performance, discriminative schemes are also incorporated to the individual nodes, making the overall approach a hybrid one (both discriminative and generative), rather than purely discriminative in SVM. Furthermore, the image representation is local and structural, instead of global or block-based.

From the TRECVID 2005 results (see Figure 1), the parts-based approach significantly improved upon the baseline, consistently for every concept. The MAP was increased by about 10%. For the "US Flag" concept, the improvement by fusing the parts-based detection with the baseline was as high as 25%, making it the best performing run. In contrast, the improvement by fusing text-based classifiers was marginal, only 2-3% in MAP. Fusion of multiple classifier instances of the same baseline model (by varying the training pools or SVM parameters) also resulted in small performance differences. This confirms that parts-based approach is powerful for detecting generic visual concepts, especially those dominated by the local attributes and topological structures.

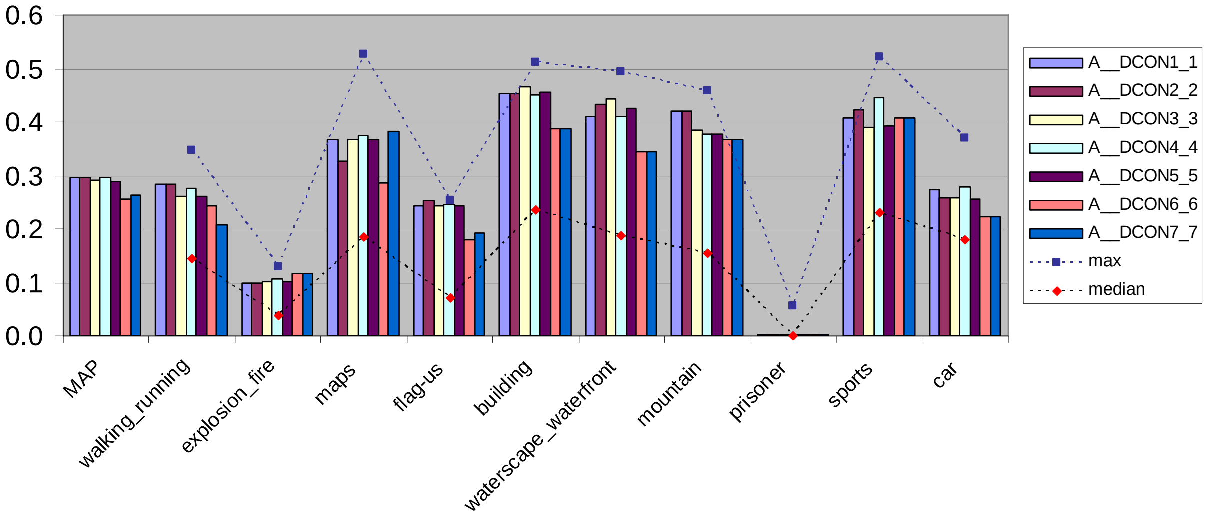

Figure 1: Performances of our submitted runs for each concept and the average performance (MAP) over the 10 concepts. The vertical axis shows the average precision (AP) values. The blue (red) points show the max (median) performance from all NIST 2005 runs. Note the parts-based detector is especially effective for the "flag-US" concept, which has strong cues from local parts as well as their topological relationships.

Search Task

For the search task, we explored several novel approaches to leveraging cues from the audio/visual streams to improve upon standard text and image-example searches in all three video search tasks. We employed "Cue-X re-ranking" [3] to discover relevant visual clusters from rough search results; "concept search" to allow text searches against concept detection results; and "near duplicate detection" [5] for finding re-used footage of events across various sources. We also apply our story segmentation framework [3,4] and share the results with the community. In the end, we find that each of these components provides significant improvement for at least some, if not all, search topics. Combinations of these new tools achieved top AP for four topics (Mahmoud Abbas, fire, boat, people/building) and good performance for an additional ten topics. We develop an analysis tool [8] to take an in-depth look at the logs from interactive runs and gain insight into the relative usefulness of each tool for each topic.

The **story segmentation** algorithm uses a process based on the information bottleneck principle and fusion of visual features and prosody features extracted from speech [3,4]. The approach emphasizes the automatic discovery of salient features and effective classification via information theory measures and was shown to be effective in the TRECVID 2004 benchmark. The biggest advantage of the proposed approach is to remove the dependence on the manual process in choosing the mid-level features and the huge labor cost involved in annotating the training corpus for training the detector of each mid-level feature. For this year's data different detectors are trained separately for each language. The performance, evaluated with TRECVID metrics (F1), is 0.52 in English, 0.87 in Arabic, and 0.84 in Chinese. The results were distributed to all active participants. In our experiments, the story segmentation was used primarily to associate text with shots. We found that story segmentation improves text search almost 100% over ASR/MT "phrase" segmentation. In the interactive system, we also enabled exploring full stories and find that a significant number of additional shots can be found this way, especially for named person topics.

Near-duplicate detection [5] uses a parts-based approach to detect duplicate scenes across various news sources. In some senses, it is very similar to content-based image retrieval, but is highly sensitive to scenes which are shown multiple times, perhaps from slightly different angles or with different graphical overlays on various different channels. It rejects image pairs where general global features are similar and retains only pairs with highly similar objects and spatial relationships. We apply near-duplicate detection in interactive search as a tool for relevance feedback. Once the searcher finds positive examples through text search or some other approach, they can look for near-duplicates of those positive shots. We have found that duplicate detection, on average, tends to lead to double the number of relevant shots found when used in addition to basic text, image, and concept searches.

In our **concept search** approach, we enable text searches against the visual content of the images. Both the text queries and subshots are represented in an intermediate concept space, containing confidences for each of the 39 concepts. The subshots are represented by the outputs of the concept detectors for each of the concepts, smoothed according to the frequencies of each concept and the reliability (performance) of each concept detector. The text queries are mapped into the concept space by measuring the semantic similarity between the terms in the query and the terms in the concept's description (via Renick's WordNet-based metric). The subshots are ranked according to their distance from the text query in concept space. This approach is applied to automatic, manual, and interactive searches with high performance in the few topics which have high-performing correlated concepts (such as "boats," "cars and roads," "military vehicles," "people with banners," and "tall buildings").

Cue-X re-ranking [3] is applied to both automatic and interactive search to re-rank the search results by balancing the visual features and pseudo-relevance labels using a framework based on the information bottleneck principle. In the automatic task, we apply the re-ranking to basic story-level text searches. Highly-ranked results are taken as pseudo-positive and lower-ranked results are pseudo-negative. The framework learns the recurrent relevant and irrelevant low-level patterns from the estimated pseudo-labels in the rough search results and reorders them according to the smoothed cluster relevance. In the automatic task, re-ranking improves MAP over the story text baseline by 15%, with particularly good results in sports and named person topics. In the interactive task, the hard negative and positive labels from the searcher can be fused with rough search scores to estimate pseudo-labels for further relevance clustering, but the improvement is somewhat more muted since interaction already focuses a great number of positives high in the search results and re-ranking further down the list has less of an effect.

The following table shows the relative contribution from each tool in our multi-modal automatic run in TRECVID 2005. Note the MAP shown below (0.114) is slightly different from our NIST official result (0.1) for the automatic run after we corrected a coding bug.

MAP	Components
0.039	Text
0.087	Text+Story
0.095	Text+Story+Anchor Removal
0.107	Text+Story+Anchor Removal +CueX Re-rank
0.111	Text+Story+Anchor Removal +CueX Re-rank +CBIR
0.114	Text+Story+Anchor Removal +CueX Re-rank +CBIR+Concept Search

Table 1. Relative Contributions of Different Search Tools (Automatic Run)

For interactive runs, we have integrated all of the above tools into a single interactive search system, including text search against ASR/MT transcripts, text search over automatically detected visual concepts (39 from LSCOM-Lite), story-based browsing, near-duplicate detection, content-based image retrieval (based on a single color feature-grid color moments), and cue-X re-ranking. Figure 2 shows the performance of our two interactive runs, compared with the max and median performance among all NIST TRECVID 2005 runs.

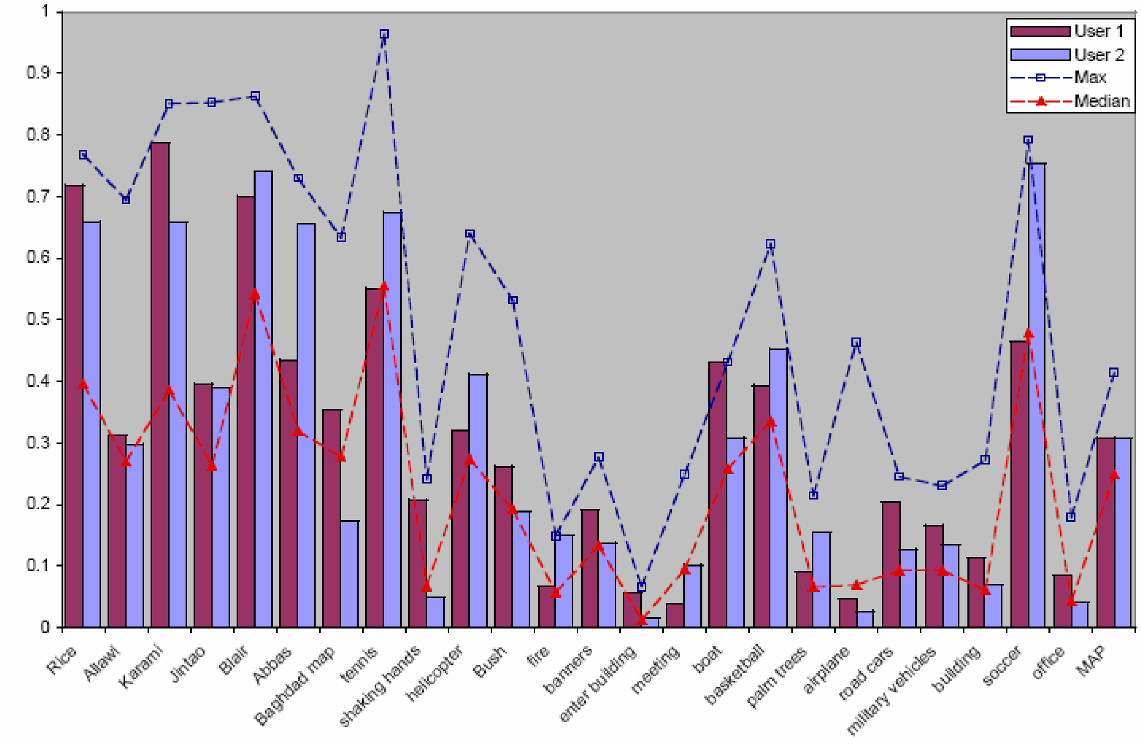

Figure 2: Performances of full interactive runs I_C_2_Columbia1_1 (User 1) and I_C_2_Columbia2_2 (User 2) for each query topic.

We have kept detailed **logs of user actions** including the queries that were issued, the browsing views that were used, and times at which positive shots were found and saved. We have developed a system for analyzing and visualizing these logs [8], which leads us to a greater understanding of the types of searches and browsing modes that are most useful for various types of queries. Looking at interactive search results, we see that we achieve the top overall performance on three topics: 160 (fire), 164 (boat), and 162 (entering building). Log analysis helps us gain a deeper of why these topics were answered so well. Figure 3 and 4 compare the percentages of true relevant results found by using each tool (near-duplicate vs. search followed by subshot browsing vs. story-based browsing) by each of two interactive searchers.

For boats and people entering buildings (in Figure 3), we find that concept search provides a strong topic baseline and duplicate detection can double the number of relevant shots found. We also found that we can achieve performance close to the best for a number of topics: 149 (Rice), 151 (Karami), 153 (Blair), 154 (Abbas), 157 (shaking hands), 161 (banners), 166 (palm trees), 168 (roads/cars), 169 (military vehicles), and 171 (soccer).

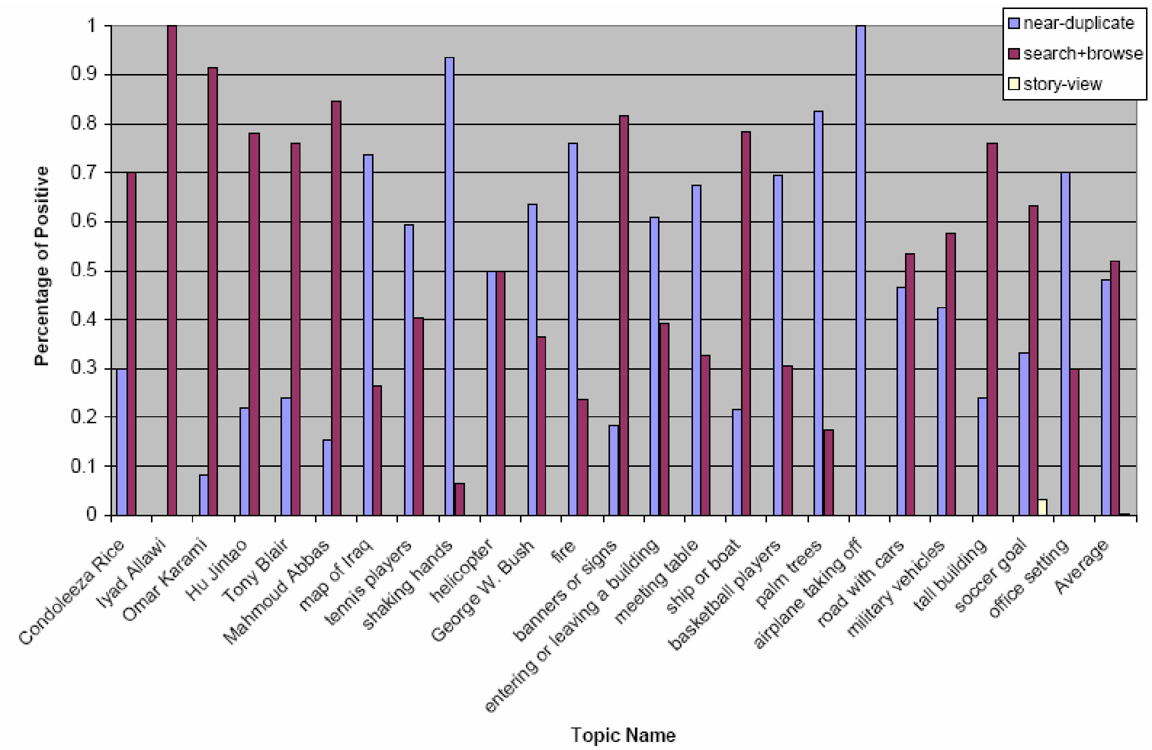

Figure 3: Percentage of true positives identified by interactive modes for user 1 for each search topic (run ID: I_C_2_ColumbiaI1_1). This user ranked best overall for topics "entering building" and "ship or boat". This user dispreferred story view and found slightly more positives with the regular search and browse method.

For fire (in Figure 4), text and content-based image search work well to find a few examples. Browsing into the stories for those found examples can double the number of relevant shots found and then using duplicate detection for all of those found shots can more than quadruple the found shots.

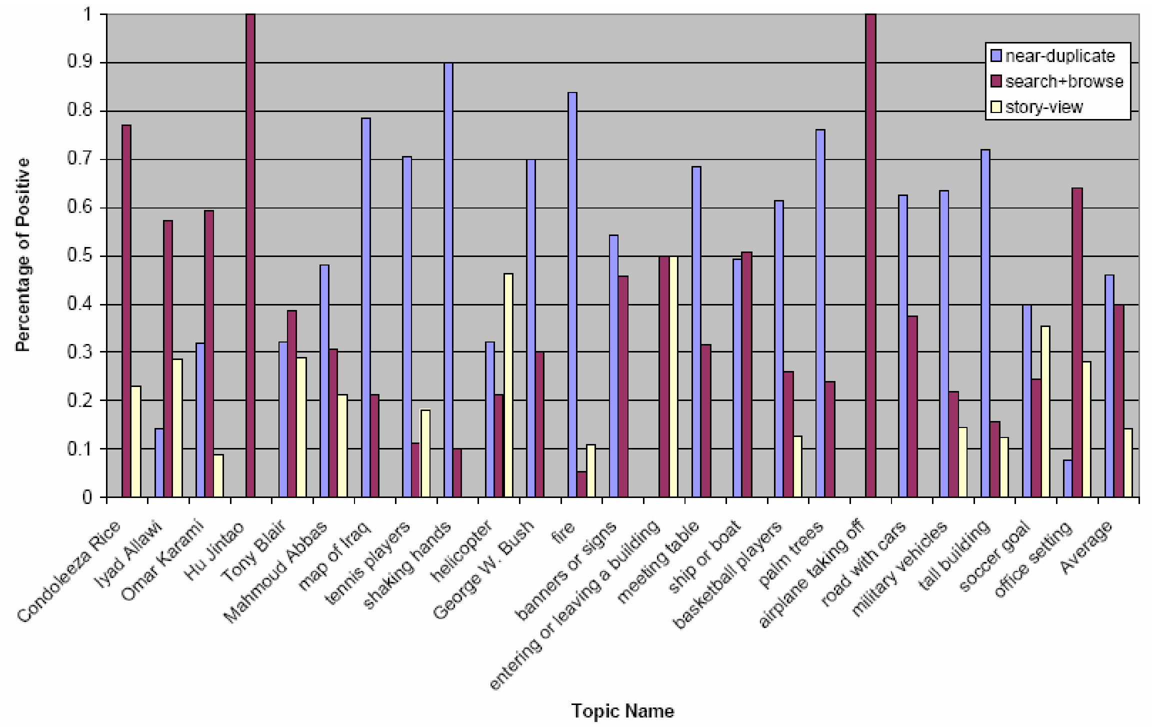

Figure 4: Percentage of true positives identified by interactive modes for User 2 for each search topic (run ID: I_C_2_ColumbiaI2_2). This user ranked best overall for topic 160 ("fire"). This user heavily used story view as well as near-duplicate search; on average the user found slightly more positives near-duplicate browsing with than with regular search and browse method.

Analyzing the logs, we can see that the formula for success for each of these topics involves finding some positives through some method and then browsing through the near-duplicates. The methods for finding the most results, though, vary by topic. For the named persons, text search and story browsing will turn up a good deal of results. Duplicate browsing will then turn up a significant number more results, usually from different sources, where the ASR/MT failed or the person's name is not mentioned. For some of the other topics (banners, palm trees, roads/cars, and military vehicles), concept search works particularly well, while for other topics (shaking hands and soccer) we have to rely on other approaches.

In addition to the above major components, we have also developed anchor shot detection [7] and query-class dependent retrieval models [6] and tested them in TRECVID 2005 search, although the relative contributions of these components have not been shown to be as significant as others mentioned earlier.

The lack of added benefit from query-class-dependent retrieval models is particularly unexpected and merits further investigation and discussion. We have concluded that their failure to improve performance is largely due to the strengths and weaknesses of each of the tools that we had available in our system. The strong tools (story-based text search and Cue-X re-ranking) performed well on many of the same query topics. The other tools (content-based image search and concept search) were generally weak and did not perform very well for any topics, including the ones which could not be answered satisfactorily with the stronger tools. For query-class-dependent models to be effective, it is necessary to have a complementary set of tools, where particular tools (or combinations of tools) perform well for unique sets of query topics.

In Figure 5, we see the break-down of the performance of our multimodal automatic search system on each of the 24 query topics. We see that the system performs quite well on the six queries for named persons and better than average on the three queries for sporting events, but quite poorly on the remaining 15 queries. The named persons and sporting events queries all respond well to story-based text search and Cue-X re-ranking, while the other queries respond fairly poorly to all of the tools (including content-based image search and concept search), so it is difficult to get any performance gain, in terms of MAP, from introducing any query-class-dependence. However, if we were to improve upon our weaker tools, it would be quite possible to better address many of the more difficult queries, or even to merit the use of different tools for sports and named person queries, and it would be helpful (and, in fact, necessary) to have a query-class-dependent approach available.

It is clear, then, that query-class-dependent models cannot provide much improvement on their own. They must be supported by a set of robust and complementary search tools which can successfully address many different types of queries.

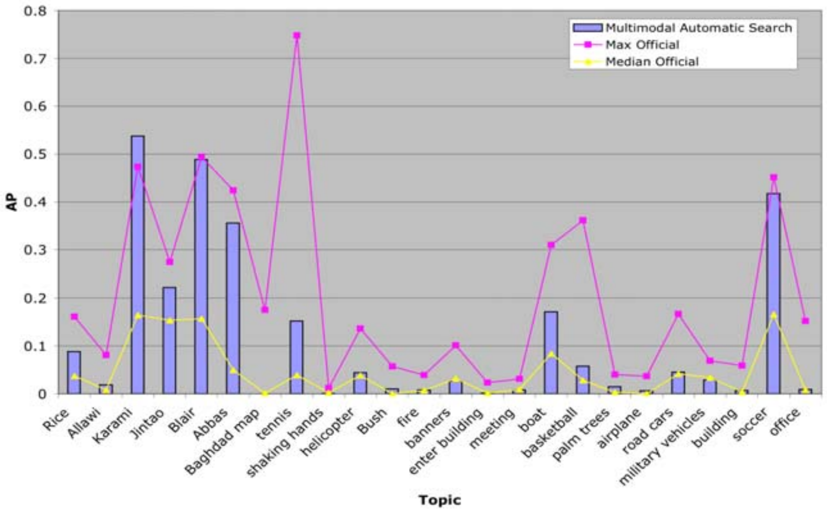

Figure 5. Results of multimodal automatic search on each query topic compared to median and max of all official runs.

Summary

We have experimented with a variety of new video search tools and found each to be quite powerful in various applications: near-duplicate detection for interactive search; Cue-X re-ranking for automatic text search; concept search for concept-related topics; and story detection for text/shot association and interactive browsing.

References:

(papers are available for download at http://www.ee.columbia.edu/dvmm)

1. Dong-Qing Zhang and Shih-Fu Chang, "Learning random attributed relational graph for part-based object detection", Columbia University ADVENT Technical Report #212-2005-6, May, 2005.
2. Arnon Amir, Janne O Argillander, Marco Berg, Shih-Fu Chang, Martin Franz, Winston Hsu, Giridharan Iyengar, John R Kender, Lyndon Kennedy, Ching-Yung Lin, Milind Naphade, Apostol (Paul) Natsev, John R. Smith, Jelena Tesic, Gang Wu, Rong Yan, Donqing Zhang, “IBM Research TRECVID-2004 Video Retrieval System**,”** In *NIST TRECVID 2004 Workshop*, Gaithersburg, MD, 2004.
3. Winston Hsu, Shih-Fu Chang, “Visual Cue Cluster Construction via Information Bottleneck Principle and Kernel Density Estimation,” In *International Conference on Content-Based Image and Video Retrieval (CIVR)*, Singapore, 2005.
4. Winston Hsu, Lyndon Kennedy, Shih-Fu Chang, Martin Franz, John R. Smith, “COLUMBIA-IBM NEWS VIDEO STORY SEGMENTATION IN TRECVID 2004,” *ADVENT Technical Report #207-2005-3 Columbia University*, 2005.
5. Dong-Qing Zhang, Shih-Fu Chang, “Detecting Image Near-Duplicate by Stochastic Attributed Relational Graph Matching with Learning,” In *ACM Multimedia*, New York City, USA, October 2004.
6. Lyndon Kennedy, Paul Natsev, and Shih-Fu Chang, “Automatic Discovery of Query Class Dependent Models for Multimodal Search,” ACM Multimedia Conference, Nov. 2005, Singapore.
7. Akira Yanagawa, Winston Hsu, and Shih-Fu Chang, “Anchor Shot Detection in TRECVID-2005 Broadcast News Videos,” Columbia ADVENT Technical Report, ##, 2005.
8. Eric Zavesky, Lyndon Kennedy, and Shih-Fu Chang, “Understanding User Strategies and Tool Performances in Video Search -- Analysis of User Action Logs,” Columbia ADVENT Technical Report, ##, 2005.