Columbia University TRECVID-2006
Video Search and High-Level Feature Extraction

Shih-Fu Chang, Winston Hsu, Wei Jiang, Lyndon Kennedy,
Dong Xu, Akira Yanagawa, and Eric Zavesky

Digital Video and Multimedia Lab
Department of Electrical Engineering
Columbia University
http://www.ee.columbia.edu/dvmm

(updated March 1st, 2007)

Description of Submitted Runs

High-level feature extraction

- **A_CL1_1:** choose the best-performing classifier for each concept from all the following runs and an event detection method.
- **A_CL2_2:** (visual-based) choose the best-performing visual-based classifier for each concept from runs A_CL4_4, A_CL5_5, and A_CL6_6.
- **A_CL3_3:** (visual-text) weighted average fusion of visual-based classifier A_CL4_4 with a text classifier.
- **A_CL4_4:** (visual-based) average fusion of A_CL5_5 with a lexicon-spatial pyramid matching approach that incorporates local feature and pyramid matching.
- **A_CL5_5:**(visual-based) average fusion of A_CL6_6 with a context-based concept fusion approach that incorporates inter-conceptual relationships to help detect individual concepts.
- **A_CL6_6:** (visual-based) average fusion of two SVM baseline classification results for each concept: (1) a fused classification result by averaging 3 single SVM classification results. Each SVM classifier uses one type of color/texture/edge features and is trained on the whole training set with 40% negative samples; (2) a single SVM classifier using color and texture trained on the whole training set (90 videos) with 20% negative samples.

Search

- **F_A_1_COLUMBIA_RR1_textbaseline_6:** text baseline system.
- **F_A_2_COLUMBIA_RR3_storyib_5:** text search using (multimodal) story boundaries and IB reranking of results (based on visual content of returned shots).
- **F_B_2_COLUMBIA_RR5_storyibcon_4:** query class dependent combination of IB reranked story-based text search and concept-based search.
- **F_B_2_COLUMBIA_RR6_viscon_3:** query class dependent combination of concept-based search and visual example-based search. Does not use speech transcripts.
- **F_B_2_COLUMBIA_RR8_textibviscon_2:** query class dependent combination of IB reranked story-based text search, concept-based search, and visual example-based search.
- **F_B_2_COLUMBIA_RR9_storyqeibteviscon_1:** query class dependent combination of IB reranked story-based search (with query expansion) fused with example-based text search, concept-based search, and visual example-based search.

Abstract

We participated in two TRECVID tasks in 2006 – "High-Level Feature Extraction" and the automatic type of "Search." Summaries of our approaches, comparative performance analysis of individual components, and insights from such analysis are presented below.

1. Task: High-Level Feature Extraction

In TRECVID 2006, we explore several novel approaches to help detect high-level concepts, including the context-based concept fusion method [13] that incorporates inter-conceptual relationships to help detect individual concepts; the lexicon-spatial pyramid matching method that adopts pyramid matching with local features, event detection with Earth Mover's distance that utilizes the information from multiple frames within one shot for concept detection [14], and text features from speech recognition and machine translation to enhance visual concept detection. In the end, we find that each of these components provides significant performance improvement for at least some, if not all, concepts.

1.1 Baseline

For the baseline detectors, we adopt a fused model by averaging two simple generic methods that use Support Vector Machines (SVM) over three visual features, namely color moments over 5x5 fixed grid partitions, Gabor texture, and edge direction histogram from the whole frame. The description of these features can be found in [1]. In the first method, three SVM classifiers are trained for each concept, one with one type of feature, using 40% negative training samples. The classification results are fused through averaging to generate a final decision. In the second method, only the color moments and Gabor texture are concatenated into a long feature vector to train SVM classifiers with 20% negative training samples. Such baseline technique has been shown competitive in past TRECVID experiments and literatures. We deliberately choose to use a generic visual-only method for concept modeling so that we may accommodate a large pool of concept models and our experiment will not be biased by any specific modeling technique.

1.2 Inter-conceptual Relationships

Contextual relationships among different semantic concepts provide important information for automatic concept detection in images/videos. For example, the presence of "outdoor" concept in an image presumably will increase the likelihood of detecting "car" and "parade". We have seen quite a few attempts in the literature exploring such contextual fusion to improve detection accuracy, but experimental results reported so far have been mixed – some concepts benefit while others degrade. This year with the much larger set of concept annotations from LSCOM [8] and MediaMill [9], TRECVID 2006 provides a ripe point for assessing the true potential of contextual fusion. Therefore, we specifically focus on evaluation of a new information theoretic method that can be used to predict whether or not a specific concept may benefit from contextual fusion with a large pool of concept models, namely 374 LSCOM concept models. Our objective is to find a judicious way of applying contextual fusion, rather than blindly using it for the whole set of concepts. Such a selective mechanism is important for optimizing the system performance in both accuracy and computational complexity.

Our prediction method takes into account both the strength of inter-conceptual relationships and the robustness of each baseline detector. A concept is predicted to be amenable to contextual fusion when its correlated concepts show strong detection power and its own detection accuracy is relatively weak. Pair-wise mutual information among concepts are used to estimate the concept correlations. Based on the TRECVID 2005 development data, our method predicts 16 out of the 39 LSCOM-lite [10] concepts will benefit from contextual fusion with the large pool of 374 concept models. Using a separate validation data set, we found the prediction has been very accurate – 13 out of the 16 predicted concepts showed significant performance gains while the remaining 3 did not show performance difference.

In TRECVID 2006 (see Figure 1), 4 out of the 16 predicted concepts are evaluated by NIST. Again, the prediction accuracy has been consistently high -- 3 of the predicted concepts, "car", "meeting" and "military-personnel" show significant improvements at about 30%, with the 4th one showing no

performance change. Such a high level of prediction accuracy is very encouraging, confirming the effectiveness of the prediction method across data from different years. It also addresses the open issue about the inconsistent effect of contextual concept fusion mentioned earlier.

Some details of the experiments are described below. We built baseline detectors for 374 of the 449 concepts included in the LSCOM annotations, excluding only the concepts with too few positive examples. For contextual fusion, we adopt an emerging framework, called Boosted Conditional Random Field (more details in [11,13]). Such a framework captures the inter-concept relationships by a Conditional Random Field. Each node in the model represents a concept and edges represent relationships between concepts. Detection scores of 374 LSCOM concepts generated by the baseline detectors are taken as input observations. Through graph learning, the detection results for each of the target concepts are refined. The graph structure, namely the joint conditional posterior probability of class labels, is iteratively learned by the well-known Real AdaBoost algorithm.

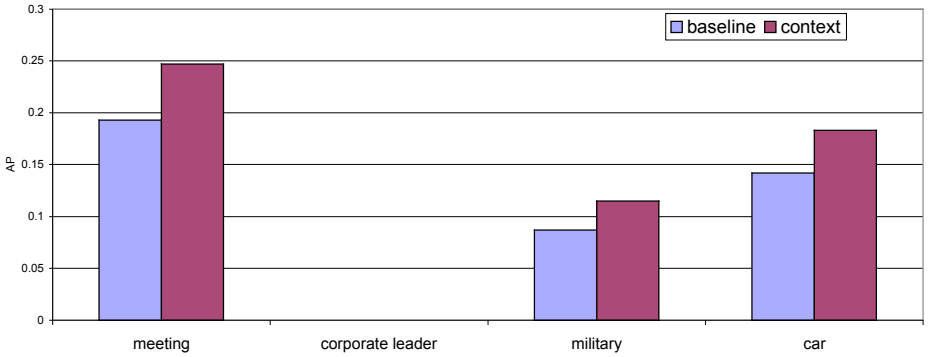

Figure 1: Performances of Context-Based Fusion by Boosted Conditional Random Field over the 4 concepts that are automatically predicted to be promising and are evaluated by NIST.

1.3 Lexicon-Spatial Pyramid Match Kernel

We generalize the idea of spatial pyramid matching [12] into a new method called lexicon-spatial pyramid matching (LSPM). Pyramid matching based on local features has been an active research topic in the computer vision community in recent years. It has been shown effective in finding correspondences between images for object/scene detection. SIFT features are extracted over each 8x8 block, based on which a bag of words image representation is adopted. By repeatedly dividing the spatial coordinates of images we compute the histogram matching of visual words at increasingly fine resolutions in the spatial domain. Similarly we repeatedly divide the visual words and calculate the spatial pyramid matching at increasingly fine resolutions in the lexicon domain. This LSPM paradigm complements our baseline method since the SIFT features capture the local attributes of object positions. The lexicon-spatial pyramid match kernels are fed into SVM classifiers for detecting individual concepts.

This LSPM approach is applied to 13 concepts, e.g., “maps”, “us-flag”, etc, which are selected since they are intuitively promising for this method. In TRECVID 2006 (see Figure 2), 6 out of these 13 concepts are evaluated, among which 3 concepts get significant performance gain, e.g., 89% for “waterscape-waterfront”. No performance degradation is observed. This confirms that the LSPM method has good potential for improving detection of TRECVID concepts, especially those dominated by local attributes.

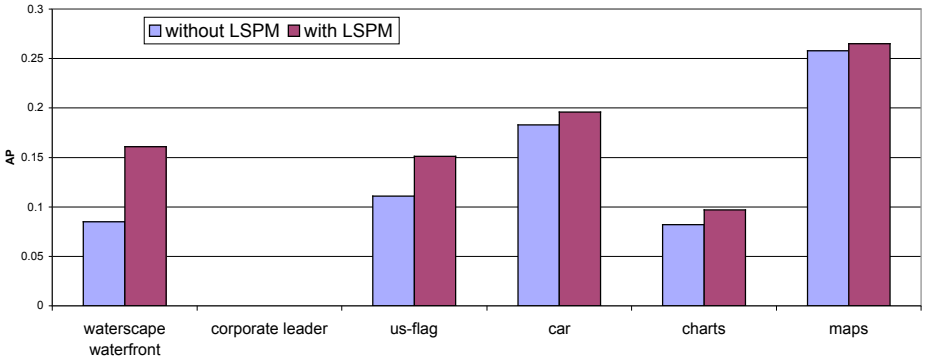

Figure 2: Performances of Lexicon-Spatial Pyramid Matching over the 6 concepts that are evaluated by NIST.

1.4 Event Detection

We explore how to utilize the information from multiple frames within one shot for event detection in contrast to the traditional key-frame based techniques [14]. Each frame is represented by a set of static concept scores, which are generated by the baseline detectors. The earth mover's distance (EMD) is adopted to calculate the distance between two shots that have different numbers of frames. Being a cross-bin dissimilarity measurement, EMD can handle distributions of variable lengths. The EMD matching scores between all video pairs (including both positive and negative samples) are then used as the kernel matrix, which is then used to train a SVM-based event detector.

The event detection method is applied to "people marching", which is related to events. In our internal test, EMD based algorithm achieves 20% relative improvement, but in TRECVID 2006 results (see Figure 3), no improvement is obtained. There are some possible reasons: (1) the data distribution of TRECVID 2005 and TRECVID 2006 is different on "people marching," EMD based algorithm may suffer from the over fitting; (2) because of the limitation of time, we only applied the EMD method to re-rank the top 5,000 subshots from the baseline detectors. It may not be feasible to achieve better performance if the top 5,000 subshots of the baseline results are relatively poor; (3) the subshot boundaries used are based on coarse estimation, instead of the true boundaries later released by NIST. Such coarse estimations are sometimes poor. For example, nearly 3% subshots (about 150) were empty because the starting time was later than the ending time.

In our follow-up work, we have expanded the initial method by introducing multi-resolution segmentation in the temporal dimension of the videos and measure the EMD-distance among videos at multiple resolutions. Such multi-resolution matching method allows for flexible temporal alignments of video frames while preserving the proximity of neighboring frames at various scales. Our experiments over ten (10) different event concepts confirmed the significant performance improvement of the EMD event detection method over the single-keyframe baseline approach [14].

1.5 Text

Text transcripts from speech recognition and machine translation associated with video shots are a powerful source of information for many visual concepts. In one of the runs, we exploit this information source by incorporating text features into our concept detection framework. In news video broadcasts there is often a great deal of asynchrony between the words being spoken by the anchor and the visual concepts appearing in the visual stream, such that the words said during a particular video shot might not be related to the content of that specific shot, but rather some adjacent shot. It is important, then, to include some mechanism for associating text from adjacent shots with a particular shot. One mechanism might be to just use a fixed window and define the text associated with a shot as the text occurring within a fixed number of shots or seconds from the shot. Our past experience with the search task has indicated that simply associating all the text within an entire story with a shot is the most effective way of dealing with this

asynchrony [4]. This makes sense since story units are semantically related to the visual content in the stream. If an anchor mentions a visual concept, it is likely to appear somewhere within the same story, but is unlikely to appear across a story boundary. We use automatically detected story boundaries, which we generated and shared with the entire community [2].

Each story is represented by a bag of words using tf-idf features, using a vocabulary of terms which appear in at least five stories in the training set. The text tokens are also stemmed using Porter's algorithm and a standard list of stop words are removed. A story is considered as positive for a given concept if one of the shots contained within that story has been annotated as positive for the concept. The story is negative otherwise. The tf-idf features and positive/negative labels are then fed into a support vector machine to learn a story-level text-based concept detector. Since there are more terms in the vocabulary than there are total training stories, some feature selection is needed to trim the tf-idf vectors down to a reasonable size. For this, we simply use the tf-idf values themselves. We take the average tf-idf score for each term over the positive examples, using only the top k terms as features. This is reasonable since tf-idf is a mechanism engineered to give weight to the terms with the most information about the content of a document. We choose the value k, along with the SVM parameters through a grid search to optimize AP over a validation set. The values are selected separately for each concept.

Text features are applied in two of the submitted runs. In A_CL3_3, text features are applied to all 20 test concepts by averaging the scores resulting from text features with the scores resulting from each of the visually-based methods. This gives a relative improvement in mean inferred average precision of 12% when compared to using visual features alone. The improvement is concentrated in a few concepts, which intuitively have a strong relationship to text. Specifically, the "explosion/fire" concept shows 40% relative improvement from text. "Sports," "weather," and "military personnel" all also demonstrate significant improvements from text, each improving approximately 25%, relative. A number of concepts, which intuitively do not have strong ties to text actually exhibit a decrease in performance by adding text. Specifically, "desert," "mountain," and "people marching" are all hurt by text. The performance change across all concepts is shown in Figure 3. Text is also used in run A_CL1_1. Here, we choose a more aggressive fusion strategy, choosing the best among our available detectors, according to a validation set, on a concept-by-concept basis. Therefore, it is possible to choose text, alone, as the submitted scores for a particular concept. This approach improves mean inferred average precision by 5.6% when compared to a similar approach based on best-of-visual detectors, only. However, this best-of-best fusion method does not perform as well as the flat, concept-independent fusion method of A_CL3_3. Clearly, there is some over-fitting at play here.

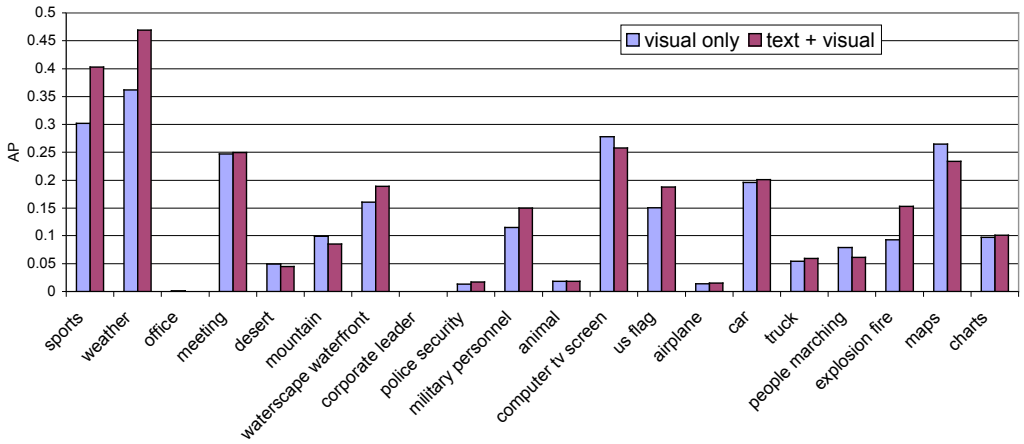

Figure 3: Performances of text-based high-level feature detection over all 20 concepts.

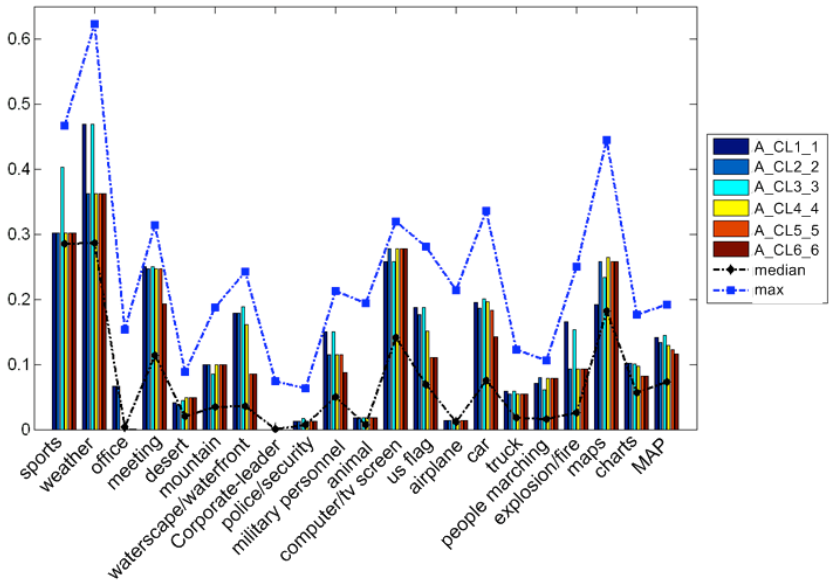

Figure 4: Performances of our submitted runs for each concept and the average performance (MAP) over the 20 concepts. The vertical axis shows the average precision (AP) value. The blue (black) points show the max (median) performance from all NIST 2006 runs.

2. Task: Search

The use of pre-trained concept detectors for semantic search over image and video collections has gained much attention in the TRECVID community in recent years. With the recent releases of large collections of concepts and annotations, such as the LSCOM [8] and MediaMill [9] sets, the TRECVID 2006 benchmark presented one of the first opportunities to evaluate the effects of using such large collections of concepts, which are orders of magnitude larger than the collections of several dozen concepts that have been available in past years. In TRECVID 2006, we specifically set out to explore the potential of using a large set of automatically detected concepts for enhancing automatic search. In particular, we explored which query topics can benefit from the use of concept detectors in the place of or in addition to generic text searches.

The concept search results were used in combination with many other search components, such as text search, query expansion, information bottleneck (IB) reranking, and visual example-based search. Results between various components were fused using a query-class-dependent fusion approach, where the weights applied to each component search method were varied depending upon the type of query. Applying all of the available components resulted in an increase in performance of 70% relative to the text baseline. The use of concept detectors in the concept-based search method gave the largest improvement of any individual method, improving 30% relative to the text story baseline.

We also examine the impact of using a large collection of concepts (374 concepts) instead of the smaller, standard set of 39 concepts. We find that this increase in concept lexicon size gives an increase of search performance of 30%, which is significant, but not proportional to the 1000% increase in lexicon size.

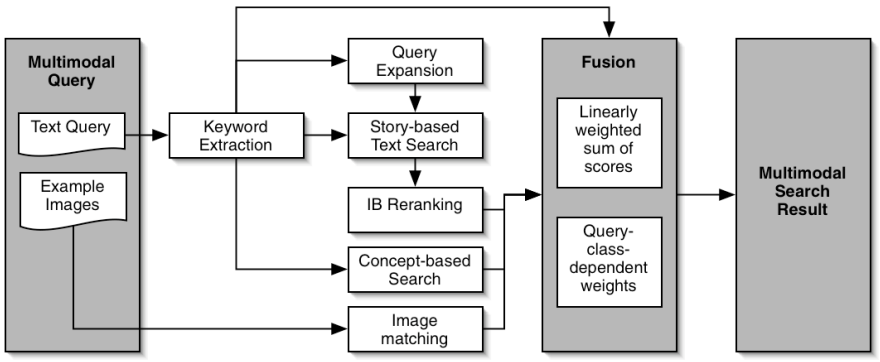

Figure 5: Architecture of multimodal video search system.

2.1 Text Search

Text transcripts from speech recognition and closed captions remain among the best resources for news video search. We exploited the speech recognition and machine translation transcripts included with the search test set in a number of ways.

2.1.1 Text Baseline

We implemented a baseline text search system which extracts keywords from the text query topics, extracting keywords like named entities and nouns, and uses those keywords for searches against the text transcripts. In this baseline system, the speech recognition transcripts are segmented into phrases. These phrase boundaries are provided automatically for the machine-translated transcripts for Arabic and Chinese programs, but there are no such boundaries for the English programs and the phrase boundaries are generated using speaker turn and timing information. When phrases matching the keywords are found, the scores are propagated to any shots occurring within either 10 seconds before or after the matched phrase. This approach performs with a MAP of .035, which ranks fourth overall when compared to other required text search baselines.

2.1.2 Story Segmentation

We ran our story segmentation system over the search set using the method developed in [2]. We used the resulting boundaries to formulate text documents in the search set. The same keyword based search method as used in the baseline was applied (using story boundaries instead of phrase boundaries) and the scores of matching documents were propagated to all shots occurring temporally within 10 seconds before or after the story. In past evaluations, we have seen much improvement from story segmentation, typically a 25% relative improvement in MAP [4]. However, in this year's evaluation, story-based text search did not seem to change the results significantly when compared to the phrase-based text search baseline. This may be due to poor quality of segmentation results caused by the existence of many programs in the search set which are not seen in the training set.

2.1.3 Query Expansion

We applied a system that performs query expansion with named entities. The system mines a large corpus of saved news broadcasts from the same timeframe as the search set to find proper names of people, organizations, and locations which are related to the initial query topic. The results show a small increase in performance from including these terms in the search. The query expansion system and its evaluation are discussed in greater detail in [7].

2.1.4 Text Example-based Search

In text example-based search, we leverage the video examples provided in the query topic. These examples have speech recognition transcripts associated with them, which can be mined to boost the performance of text keyword search. For a given query topic, we train support vector machine classifiers, using the provided video examples as positive examples and randomly selected news stories as pseudo-negative

examples. The features associated with the examples are simply tf-idf features, selected to reflect the most frequent terms in the positive examples (much like in our text-based concept detection system discussed earlier). Text features are extracted from entire stories and shots are associated with the story in which they temporally occur. 15 different SVM models are learned, using the same examples as positive training examples and changing which documents are selected as pseudo-negatives. Each of the learned models are applied to the stories in the search set and the various resulting scores are averaged to give a final example-based score. We averaged the scores resulting from the story-based text search with the example-based text search. This gives a 25% relative increase in MAP and performs particularly well for the three named persons queries which have examples in the query topic ("Dick Cheney", "George W. Bush", and "Condoleezza Rice"), improving 75% relative for those three queries.

2.1.5 Information Bottleneck Reranking

Information bottleneck (IB) reranking is a visual-based reordering which we apply as a post-processing step to our various text-based searches [3]. The IB reranking framework processes the results from a text search to find clusters of visually recurrent patterns, which have high mutual information with highly relevant documents. The top results from the text search are then reordered to push up relevant recurrent patterns and push down irrelevant ones. The IB reranking framework gives 10% relative increase in MAP, when applied to the results of story-based text search.

2.2 Concept-based Search

Our concept-based search system leverages a large number of pre-trained concept detectors used to give a mid-level semantic representation of each shot in the search set. Text-based keyword searches can then be mapped through this concept space to find shots with visual content that is semantically related to the search query.

We built automatic detectors for 374 of the 449 concepts included in the LSCOM annotations, excluding only the concepts with too few positive examples. The automatic detectors adopt the same approach as that used in the first fused model in our visual-based baseline detectors, namely Run A_CL6_6 listed at the beginning of the paper. The detectors are built with a simple baseline method using support vector machines (SVMs) over three visual features: color moments on a 5-by-5 grid, Gabor textures over the whole image, and an edge direction histogram. This method has been used by several participant groups in TRECVID and has been shown to achieve reasonable success over many concepts. We deliberately choose to use a generic visual-only method for concept modeling so that our experiment will not be biased by any specific modeling technique. The baseline concept detection method was submitted to the high level feature evaluation this year and performed with a mean inferred average precision of 0.116 over the 20 evaluated concepts. The 374 baseline visual detection models are available for download, including the SVM models, the visual features and classification results of TRECVID 2005 development/test sets, and TRECVID 2006 test set. We intend to extract and share the same data for future TRECVID evaluation data sets. Details about the models and the download process can be found in [15].

Our concept-based search method applied each of the 374 concept detectors to each shot in the search test set. Incoming text query topics were then processed to find concepts in the lexicon which were related to the query. A weighted average of the detection scores for the matched concepts was then used to score and rank the shots in the search test set. The weights of the concept detection scores are determined by two factors: (1) the quality of the text match between query terms and concept descriptions and (2) the reliability of the concept detector. For the text-to-query mapping, we only allow direct matches between queries and concepts, so weights due to this first factor are binary, either one or zero. The second factor, the detector reliability can be estimated as the average precision of the detector over a validation set. If the average precision is high, the detector is reliable and we can treat the provided score as a good measure of the likelihood of a concept being present in the shot. If the average precision is low, the detector is unreliable, so the provided score doesn't give us much more information about the presence of a concept in the shot, beyond the prior probability (or the frequency of the concept in the development set). Therefore, we take the absolute value of the average precision (between zero and one) and use it as a smoothing factor between the concept detection score and the prior probability of the concept presence. In queries with

multiple matching concepts, this prevents unreliable concept detectors from seriously degrading performance, while in queries matching only one concept, the ranked list provided by the detector is preserved, regardless of reliability.

In essence, the method indexed the shots in the search set based on the visual content alone, without the speech recognition transcripts, and allowed for search given only text queries, without requiring image examples. Both of these aspects are very attractive for enabling fast, semantic access to video databases.

We found much support for the utility of concept-based search in the TRECVID 2006 set of query topics. Of the 24 total query topics, 17 had direct matches to related concepts in the LSCOM lexicon. Of the 7 remaining topics, 4 were for named persons, which were purposely excluded from the LSCOM lexicon, and only 3 were for objects or concepts which were truly outside of the lexicon. We also found that among the 17 concept-related queries, there were on average 2.5 matched concepts per query.

The concept-based search method was used in combination with the text search baseline, along with several other basic components, such as story segmentation, query expansion, and visual example-based search, with each of the components applied differently for various classes of queries. In total, all of the components used together provided an improvement of 70% relative to the text baseline. The concept-based search accounted for the largest individual portion of this improvement, improving 30% relative over a story-based text search. The degree of improvement is particularly large for a few queries with strong concept detectors, such as "Find shots of one or more soccer goalposts" (140% relative improvement), "Find shots with soldiers of police escorting a prisoner" (70% improvement), "Find shots of soldiers or police with weapons and vehicles" (40% improvement), or "Find shots of tall buildings" (many fold increase). These results confirm that a large lexicon of pre-trained concept detectors can significantly enable content-based video search, particularly when the concept detectors are strong.

2.3 Visual Example-based Search

Example videos and images are used in a query-by-example method for visual example-based search. We apply a framework similar to the method proposed in [6]. The query videos and images are represented using 5x5 grid color moments, Gabor texture, and an edge direction histogram. In each of these feature spaces, test images are scored based on their Euclidean distance from the query images. SVM models are also learned using the query images as positive examples and randomly sampled images as pseudo-negatives, learning 10 models with the same positives, but varied pseudo-negatives. The final SVM score is produced by averaging the scores from the various models. All scores from Euclidean distance and SVM models in various feature spaces are then averaged, to give a final fused visual example-based search score. This method improved 14% relative in MAP when fused with text search and concept-based search. It is especially powerful for sports-related queries, improving 24% relative for the "soccer" query topic.

2.4 Query-class Dependent Fusion

The various search tools are applied for any given search query and combined through a weighted combination of scores. The weights for the various methods, though, are different depending upon the type of the query, using a query-class-dependent model [5]. Through examination of the performance of our tools over the TRECVID 2005 queries, we decided upon a set of five query classes: "named entity," "sports," "concept," "named entity with concept," and "general."

The **"named entity"** class is triggered whenever a named entity is detected in the text query topic. The four named person queries, "Dick Cheney," "Saddam Hussein," "George W. Bush," and "Condoleezza Rice," were all classified into this class. The class relies almost entirely on text search and gives a small amount of weight to visual example-based search.

The **"sports"** class is triggered whenever a term from a predefined list of sports terms is found. Only the "soccer" query was classified as part of this class. The class relies heavily on visual example-based search, but is also helped by text and concept-based searches as well.

The **"concept"** class is triggered whenever a query contains a term which matches a concept in the lexicon. By far, the most query topics were classified into this class, 16 of the total 24, such as "scenes with snow," "tall buildings," and "demonstration or protest." The class relies heavily on concept-based search, with a small amount of weight given to text and visual example-based methods.

The **"named entity with concept"** class is applied when the query topic contains both named entities and concept-related terms. The "Saddam Hussein with someone's face" and "George W. Bush walking" queries might have been most appropriate for this class; however, both were classified into the named entity class due to errors in the classification algorithm. This class gives equal weight to both text search and concept-based search.

The **"general"** class is essentially a catch-all for all query topics which do not fall within any of the first four classes. In this year's set of query topics, the "people in uniform," "people reading newspapers," and "person with 10 books" fell into this class. The class gives equal weight to both text and visual-example based search methods.

2.5 Search Experiment: Evaluating the Impact of 374 Visual Concept Detectors

One of the most unique aspects of our video search system was the usage of such a large collection of visual concept detectors. We conducted some experiments and analysis to evaluate the effects of using 374 concepts as opposed to the standard LSCOM-Lite [10] set of 39 concept detectors. In effect, we are increasing the size of concept lexicon ten fold over the conventionally-sized lexicon and we would like to investigate the influence that this change has on search performance and the number of queries which can leverage concept detection scores. We conduct our experiments by running two different versions of the concept-based search system. In one version, we simply use the concept-based search method described above. In the other version, we take the components from the first version and simply limit the range of possible matched concepts to only be the 39 LSCOM-Lite concepts, instead of the full 374-concept set.

We find that increasing the concept lexicon from 39 concepts to 374 concepts increases the mean average precision (MAP) of concept-based search from .0191 to .0244, a relative change of about 30%. This is a significant improvement; however, it is interesting that it is not proportional in magnitude to the additional cost of increasing the concept lexicon by ten fold. Further investigation shows that the 39 concepts are able to be used in 12 of the 24 query topics with about 1.8 concepts being useful for each of those matched queries. In comparison, the 374 concepts are able to be used in 17 of the 24 queries with about 2.5 concepts being useful for each matched query. Therefore, it seems that the increase in the size of the concept lexicon gives increases in query coverage and concept usage in the neighborhood of 40%-50%, for this set of query topics.

Two other factors are also likely limiters on the benefits seen from increasing the size of the concept lexicon: (1) the design of the concept lexicons and (2) the design of the query topic set. The first factor (the design of the concept lexicons) implies that the 39 LSCOM-Lite concepts are not simply a random subset of the 374 LSCOM concepts. Instead, they are a carefully selected group, designed by selecting the concepts which would be most practical for video indexing, in terms of their utility for search, the anticipated quality of automatic detectors, and their frequency in the data set. Indeed, we can observe that a lot of these design goals are shown to be true in the data. For example, we compare the frequency of concepts in each set over the development data. The 374 LSCOM concepts have, on average, 1200 positive training examples per concept, while the 39 LSCOM-Lite concepts have, on average, 5000 positive training examples per concept. Another way to look at the data is that more than half of the 374 LSCOM concepts are less frequent than the least frequent LSCOM-Lite concept. Such a disparity in availability of training data is expected to lead to a similar disparity in concept detector performances. We evaluate the concept detectors over a validation subset from the TRECVID 2005 development set and find that this detection disparity is also present. The 374 LSCOM concepts have a MAP of 0.26, while the 39 LSCOM-Lite concepts have a MAP of 0.39. In fact, roughly a quarter of the 374 LSCOM concepts have average precision values very close to zero. The second factor (the design of the query topic set) is due to the

influence that the concept lexicons have on the selection of query topics by NIST. An important goal of the TRECVID benchmark at large is to evaluate the utility of pre-trained concept detectors in video search. To achieve this goal, the benchmark organizers include many queries which should theoretically be well-suited for using concept detection results; however, the LSCOM-Lite set of 39 concepts is the standard set used in TRECVID and our larger set is non-standard, so the queries are designed to make use of the 39-concept set, without much regard to the content of our 374-concept set. This skews the evaluation queries in a way that is important to extending the opportunity to use concepts in search to as many participating groups as possible, but also potentially limits the impact that a large-scale concept lexicon might have.

2.6 Search Experiment: Query-Class-Dependent Multimodal Fusion

Our search system is designed to combine each of the core component search methods (text search, concept-based search, and visual example-based search) through a weighted sum, where the weights of each method are determined by the class of the query. We have conducted some experiments and analysis to measure the effects of each of the component methods on search performance, as well as the performance of class-dependent fusion.

In Figure 6, we see the performance of each of our component search tools, as well as the fusion of many subsets of individual tools and all of the official TRECVID 2006 automatic search submissions. We can see that using all of our available tools in combination improves over the text search baseline by 70%. The largest individual contributor to this improvement over the baseline, is the concept-based search, which provides for a 30% relative boost in mean average precision.

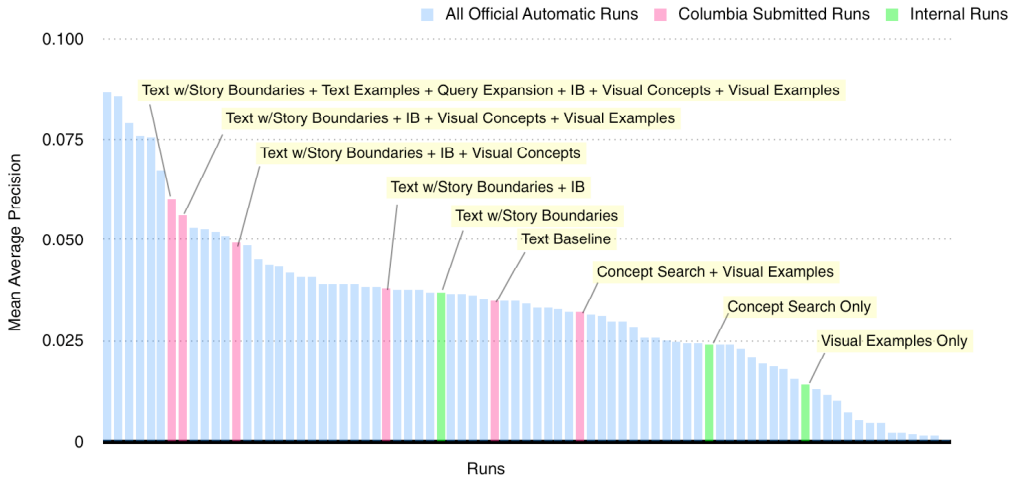

Figure 6: Performances of all Columbia automatic runs and all official automatic search submissions.

	Text	*IB-rerank*	*Concept*	*Visual*	*Fused*	Change
Building	0.04	0.03	0.30	0.10	0.33	**+10%**
Condi Rice	0.19	0.20	0.00	0.01	0.20	**+0%**
Soccer	0.34	0.42	0.58	0.42	0.83	**+43%**
Snow	0.19	0.29	0.33	0.03	0.80	**+143%**
All (2006)	**0.08**	**0.09**	**0.10**	**0.07**	**0.19**	**+90%**

Table 1: Precision of top 100 results for component and fused searches for a few example query topics. [It should be noted that the metric used (precision of top 100) is different from the standard evaluation metric (average precision). In this query topic set, average precision is typically low (due to low recall). Here, we adopt precision at 100 for its higher absolute values, which better highlight the differences between methods].

In Table 1, we see a breakdown in performance across each individual component over a few exemplary query topics. We see that the query-class-dependent approach is consistently able to select a fusion across

individual search methods that provide consistent improvement. At worst, the fusion preserves the performance of the top-performing individual method (like the Text/IB method for named person queries, such as Condoleezza Rice), and at best provides a large improvement over the top-performing individual method (like the concept-based method for concept-type queries, such as building or snow). Again, on average, the fusion across search methods provides significant gains.

2.7 Search Analysis and Conclusions

In this year's TRECVID evaluation, we primarily explored the use of 374 pre-trained concept detectors for semantic search. This concept-based search method was used in combination with a number of other methods, such as text search and visual-example based search, with the weights for each method varying depending upon the class of the query topic. All the methods combined showed an improvement of 70% relative to the text search baseline with the bulk of the improvement coming from the concept-based search method, which applies to 17 out of the 24 topics and yields a 30% relative improvement over text search alone. We also evaluated the impact of using a large collection of pre-trained concept detectors (374 concepts) instead of a smaller, standard set (39 concepts) and find that the increase in concept lexicon size provides an increase in search performance of 30%, which is significant, though not proportional to the 1000% increase in lexicon size.

3. References

[1] Akira Yanagawa, Winston Hsu, Shih-Fu Chang, "Brief Descriptions of Visual Features for Baseline TRECVID Concept Detectors", ADVENT Technical Report #219-2006-5 Columbia University, July 2006.

[2] 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, Singapore, 2005.

[3] Winston Hsu, Lyndon Kennedy, Shih-Fu Chang, "Video Search Reranking via Information Bottleneck Principle", In ACM Multimedia, Santa Barbara, CA, USA, 2006.

[4] Shih-Fu Chang, Winston Hsu, Lyndon Kennedy, Lexing Xie, Akira Yanagawa, Eric Zavesky, Dongqing Zhang, "Columbia University TRECVID-2005 Video Search and High-Level Feature Extraction," In NIST TRECVID Workshop, Gaithersburg, MD, November 2005.

[5] Lyndon Kennedy, Apostol (Paul) Natsev, Shih-Fu Chang, "Automatic Discovery of Query Class Dependent Models for Multimodal Search", In ACM Multimedia, Singapore, November 2005.

[6] Apostol Natsev, Milind R. Naphade, and Jelena Tesic, "Learning the semantics of multimedia queries and concepts from a small number of examples," in ACM Multimedia, Singapore, 2005, pp. 598–607.

[7] Zhu Liu, David Gibbon, Eric Zavesky, Behzad Shahraray, Patrick Haffner, "AT&T RESEARCH AT TRECVID 2006." Notebook Paper. NIST TRECVID Workshop, Gaithersburg, MD, November 2006.

[8] LSCOM Lexicon Definitions and Annotations Version 1.0, DTO Challenge Workshop on Large Scale Concept Ontology for Multimedia, Columbia University ADVENT Technical Report #217-2006-3, March 2006.

[9] Cees G.M. Snoek, Marcel Worring, Jan C. van Gemert, Jan-Mark Geusebroek, and Arnold W.M. Smeulders. The Challenge Problem for Automated Detection of 101 Semantic Concepts in Multimedia. In Proceedings of ACM Multimedia, Santa Barbara, USA, October 2006.

[10] M. Naphade, L. Kennedy, J. Kender, S.-F. Chang, J. Smith, P. Over, A. Hauptmann (2005). "A Light Scale Concept Ontology for Multimedia Understanding for TRECVID 2005." IBM Research Report RC23612 (W0505-104), May, 2005.

[11] Wei Jiang, Shih-Fu Chang, Alexander C. Loui, "Active Context-based concept fusion with partial user labels," In IEEE International Conference on Image Processing (ICIP 06), Atlanta, GA, USA, 2006.

[12] S. Lazebnik, C. Schmid, J. Ponce, "Beyond bags features: spatial pyramid matching for recognizing natural scene categories", IEEE Intl' Conf. on Computer Vision and Pattern Recognition, New York, 2006.

[13] W. Jiang, S.-F. Chang, and A. C. Loui. Context-based Concept Fusion with Boosted Conditional Random Fields. In IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Hawaii, USA, April 2007.

[14] D. Xu and S.-F. Chang, "Visual Event Recognition in News Video using Kernel Methods with Multi-Level Temporal Alignment," submitted for publication.

[15] A. Yanagawa, S.-F. Chang, and W. Hsu, "Columbia University's Baseline Detectors for 374 LSCOM Semantic Visual Concepts," , Columbia University ADVENT Technical Report #222-2006-8, March 2007. (download URL: www.ee.columbia.edu/dvmm/Columbia-374)