A Reranking Approach for Context-based Concept Fusion in Video Indexing and Retrieval

Lyndon S. Kennedy
Dept. of Electrical Engineering
Columbia University
New York, NY 10027
lyndon@ee.columbia.edu

Shih-Fu Chang
Dept. of Electrical Engineering
Columbia University
New York, NY 10027
sfchang@ee.columbia.edu

ABSTRACT

We propose to incorporate hundreds of pre-trained concept detectors to provide contextual information for improving the performance of multimodal video search. The approach takes initial search results from established video search methods (which typically are conservative in usage of concept detectors) and mines these results to discover and leverage co-occurrence patterns with detection results for hundreds of other concepts, thereby refining and reranking the initial video search result. We test the method on TRECVID 2005 and 2006 automatic video search tasks and find improvements in mean average precision (MAP) of 15%-30%. We also find that the method is adept at discovering contextual relationships that are unique to news stories occurring in the search set, which would be difficult or impossible to discover even if external training data were available.

Categories and Subject Descriptors

H.3.1 [**Information Search and Retrieval**]: Content Analysis and Indexing

General Terms

Algorithms, Performance, Experimentation

Keywords

concept detection, video search, context fusion, reranking

1. INTRODUCTION

Semantic indexing and retrieval over multimedia databases has been the focus of considerable interest and research over the past years. This interest has been fueled largely by the standardized test sets and benchmarks available through the NIST TRECVID video retrieval evaluation [1]. The primary thrust of TRECVID-based research has been in two main areas. The first is *concept detection*, where the goal is to automatically annotate video shots with visual concepts (usually objects, locations, or people), given a pre-defined lexicon of concepts and a sufficient number of annotated examples for building supervised detection models. The second is *search*, where the goal is to find shots related to a query of text keywords and example images or video clips, which is unknown during system development and can be for any arbitrary semantic scene, object, or person.

It has been noted that these two tasks are actually two extreme scenarios of a single task with a unified goal of finding shots matching some semantic information need [15]. Concept detection is the extremely supervised case, where thousands of examples are available for training models. On the other hand, search, is the extremely *un*supervised case, where only a few examples are given. The implications are that concept detection is labor intensive but is likely to provide accurate detection, while search requires little human input, but consequently has less predictable behavior.

One promising new direction is to utilize results from concept detection to aide in search, thereby leveraging focused human labor on a finite concept lexicon to help answer and refine infinitely many search queries [5, 3, 9]. For example, given a lexicon of a few generic concepts, such as "boat," "water," "outdoors," "sports," "government leader" or "woman," then, an incoming search query, like "Find shots of boats," could be handled by simply returning the shots for the pre-trained boat detector. Likewise, a search query like "Find shots of Condoleezza Rice," could be handled by searching against the speech recognition transcript to find occurrences of Condoleezza Rice's name, but also by giving positive weight to shots which are postive for "government leader" and "woman" and negative for "sports." From this standpoint, we might conjecture that our previous use of only the "boat" concept to answer the "Find boats" query is somewhat naive. Shouldn't there also be "water" in scenes with boats? Shouldn't they be "outdoors," too?

This notion of *context fusion*, or the use of peripherally related concepts to refine detection of semantic topics, has been explored in prior work for use in concept detection [5, 12, 16, 18]. The nature of concept detection makes it reasonable to discover related concepts through mining ground truth annotations for co-occurrences among concepts and training models on those interactions. The conclusion in previous works has been that this approach provides real, though small, improvements in detection accuracy. In our prior work [11], statistical measures based on mutual information and detector performance are also used to predict

CIVR'07, July 9–11, 2007, Amsterdam, The Netherlands.

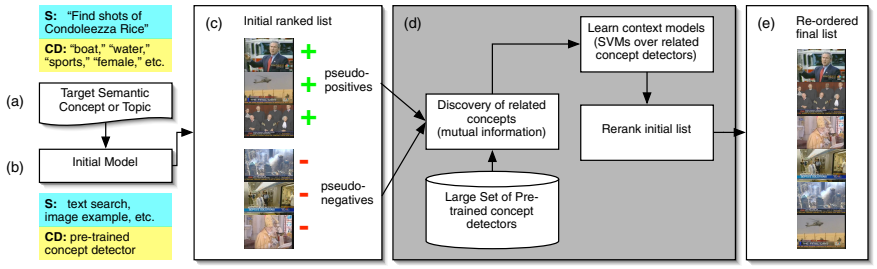

Figure 1: Architecture of proposed framework for context fusion in concept detection and search. Given a set of pre-trained concept detectors, related concepts are discovered through measuring the mutual information between their resulting concept scores and pseudo-labels given by an initial detection/search model result, such as a baseline detector or text search. Pseudo-labels and discovered related concepts are leveraged to re-order and refine the initial detection result. Examples for generic parts in (a) and (b) shown for concept detection (CD) and search (S).

the subset of concepts that may benefit from context fusion.

Using peripherally related concepts for search, on the other hand, has not yet seen such wide-spread investigation. Early approaches have included engineering intuitive filtering approaches, such as removing shots with an "anchor person" present or positively weighting the "face" concept in searches for named persons, giving small improvements [6, 8]. More recent work has included the angle of directly matching query keywords with concepts in the lexicon (like in our "Find shots of boats" / "boat" concept example) [5, 3]. However, deeper relationships to peripheral concepts are difficult to uncover, particularly in search, where the details of the searched concept are unknown to the system, due to the lack of examples provided by the human user [6, 3, 18].

In this work, we propose a framework for automatically discovering and leveraging peripherally related concepts. The approach is largely unsupervised and can therefore be applied equally well to context fusion on both concept detection and video search tasks. The approach performs surprisingly well compared to fully-supervised context fusion approaches for concept detection, with relative improvements of 7% in terms of mean average precision (MAP). The approach is also shown to give improvements of 15%-30% in MAP on video search tasks. Much of the improvements are drawn from queries for named persons or sports, while the impact on the concept-related class (which encompasses the majority of queries) is around 12%-15%.

1.1 Proposed Approach

The proposed framework in Figure 1 relies on *reranking* to automatically discover related semantic concepts and incorporate their detection scores to refine the results for concept detection or search. *Reranking* is unique in that it takes a ranked list of results from some initial search result or concept model as an approximation of the ideal semantics of the target. This initial list can then be mined to discover and leverage related concepts. Reranking is beneficial when compared to the alternative, which would require ground truth to learn context fusion models. This necessary ground truth is generally unknown in search applications.

In Figure 1, we see the flow of the reranking-based context fusion framework. When conducting concept detection or search, we need to start with a target semantic concept or topic (shown in Figure 1a), which can be a specific concept in a concept lexicon or an incoming search query. From this target, it is typically feasible to build an initial model, shown in Figure 1b. In concept detection, an initial model would be a supervised concept detector, discussed further in Section 2. In search, the initial model can be the results of a search method, such as text search, concept-based search, or example-based image matching, all of which are discussed in greater detail in Section 3. Regardless of how the initial model is obtained, we can take the results that it produces and assume that high-scoring shots are positive, while lower-scoring shots are negative, thereby creating pseudo-labels for the data in the search set, like in Figure 1c. These pseudo-labels can then be applied to discover and leverage related concepts to refine the initial results.

This approach to conducting context fusion bears much resemblance to earlier works; however, a key difference lies in the fact that this reranking method requires no *a priori* knowledge of the target semantic concept and its ground truth relationships with other concepts. We will find that despite this lack of supervision, the reranking approach performs comparably to supervised approaches.

1.2 Related Work

1.2.1 Context Fusion

As we have seen, it is intuitive that knowledge of the detection results for a large lexicon of concepts can be useful for refining the detection of individual concepts. This context-based concept fusion approach has been explored and exploited in several prior works. The Discriminative Model Fusion (DMF) method [16] generates a model vector based on the detection score of the individual detectors and an SVM is then trained to refine the detection of the original concepts. In [11], the DMF model is modified and extended to incorporate active labeling from a user. These early approaches suffer from limitations in the concept lexicons, which are typically only on the order of a few dozen concepts. The results typically show improvement on only some of the concepts, while degrading others, requiring techniques for predicting when the context fusion will succeed.

The recent releases of much larger concept lexicons [2, 17], which contain hundreds of concepts, has renewed interest in context-based concept fusion. This is explored further in [18], where hundreds of high-level features, which can be difficult to detect, are modeled using cues resulting from more-easily-detected mid-level features, such as "sky" or "person" or "outdoors." Another work in context fusion with large lexicons [12] uses an approach called a Boosted Conditional Random Field (BCRF). This framework captures the contextual relationships by a Conditional Random

Field, where each node is a concept and the edges are the relationships. Detection scores of 374 concepts generated by a baseline detection system are taken as input observations. Through graph learning, the detection results for each of the target concepts are refined.

These approaches are fully supervised and require explicit knowledge of the target semantic concept and ground-truth labels in order to discover relationships with other concepts. While this constraint is fine for concept detection, where many labels are available, it is unclear how these approaches could cover the unsupervised conditions in search.

Work in using concept detectors for context in search has been much more limited. Basic applications have included methods of filtering out or boosting certain concept types depending on the type of query, such as weighting the "face" concept for searches for people, or a "sports" concept for sports queries, or filtering out the "anchor" concept across all searches [5, 8]. These approaches provide small gains in improvement and are very difficult to scale up to effectively utilize large lexicons of hundreds of concepts.

The primary successes in leveraging concept lexicons for search have been in using direct matches between query keywords and concept descriptions to find a few concepts directly related to the query. This approach suffers from limitations in the breadth of concepts that can be applied to a query. Some empirical research suggests that it may be highly beneficial to incorporate many peripherally related concepts instead of just a few concepts that are tightly related [13]. Some approaches have incorporated lexical relationships between query terms and concepts through knowledgebases; however, these relationships are often shaky, sometimes degrading search performance by uncovering relationships that are ill-suited for video retrieval [6, 3, 18]. For example, a knowledge network might tell you that "boat" is related to "airplane" since both are types of vehicles; however, in natural images, these concepts are rarely visually co-occurrent, since they occur in very different settings. This approach is explored in [6] and [3] for concept lexicons of only a few dozen concepts, showing small improvement over direct matches, alone. This is likely due to the sparse and orthogonal concept space. In [18], the approach is applied using a lexicon of several hundred concepts, and the indirect ontology-based mapping is shown to seriously degrade search performance, when compared to direct matching, likely due to the increased risk of false selection of concepts. Our experience confirms this effect on large-lexicon context fusion for search, therefore, to use large lexicons, it is necessary to engineer methods that uncover the visual co-occurrence of related concepts with the target instead of less meaningful lexical relationships.

1.2.2 Reranking

Reranking is rooted in pseudo-relevance feedback (PRFB) for text search [4]. In PRFB, an initial text search is conducted over text documents and the top-ranked documents are assumed to be true, or pseudo-positives. Additional terms discovered in the pseudo-positive documents are then added to the query and the search is re-run, presumably providing greater clarity of the semantic target and refining the search results. An extension of PRFB from the text domain to video search is to simply apply the method to text searches over text modalities from video sources (such as the speech recognition transcripts) [7]. Further extensions have used text search to obtain psuedo-*negative* examples from video collections and used those results for example-based search with user-provided positive images [20].

In [10], the authors extract *both* pseudo-positives and pseudo-negative image examples from text search results, requiring no user-provided examples. It is found that for many types of queries, this reranking approach outperforms approaches requiring the input of example images from the user. The intuition is that many relationships between search topics and peripheral concepts change according to the news stories in the time frame of the search set and such relationships may not be present in the provided example images or external training data. Reranking can discover the salient relationships that are present in the the data to be searched.

These applications of PRFB and reranking, however, are all applied to low-level features, such as text token frequencies or color and texture image features. In a parallel work [19], the use of large sets of concept detectors to uncover contextual cues for refining search results is explored, using initial query results as a hypothesis and re-ordering those results through a weighted summation of 75 pre-trained concept detectors. The reranking mechanism used is called "probabilistic local context analysis," and functions by assuming the top-returned results are positive, while others are negative, and treats the weights on pre-defined concept detectors as latent variables to be learned. The approaches and experiments used in this paper and [19] are structured similarly. Our proposed approach includes two steps: the selection of relevant concepts and the construction of discriminative classifiers to rerank the initial search results. This is significantly different from the generative model used in [19], where latent aspects involving weighted sums of concepts are found. One potential drawback is the lack of clear semantics associated with each latent aspect, unlike the explicit relations discovered in our approach. Furthermore, discriminative classifiers have been found to be more effective than generative models in context fusion and reranking.

1.3 Outline

The remainder of the paper is as follows. In Sections 2 and 3, we discuss the concept detection and search systems used in our context fusion experiments. In Section 4, we discuss the context fusion approach. Section 5 gives the details for experiments and analysis and conclusions are in Section 6.

2. CONCEPT DETECTION

In concept detection, the objective is to build automatic detectors for arbitrary visual concepts, given a large set (on the order of hundreds or thousands) of ground-truth labels for the concept. To achieve this, we applied a generic visual-only framework which can be applied to any concept without any specific tuning. For simplicity, we choose to represent shots as single still keyframe images; however, in principle, more involved temporal models could be applied for event-based concepts. The concept model is built using SVMs over three visual features: color moments on a 5-by-5 grid, Gabor textures over the whole image, and an edge direction histogram. The resulting scores of each of the three SVM models over the test keyframe images are then averaged to give a fused concept detection score. We have found the accuracy of such baseline detectors satisfactory over a large set of concepts [5], especially for scenes and objects. The concept detection framework is discussed more in [21, 5].

We apply the concept detection framework for 374 of the 449 concepts from the LSCOM [2] annotations, excluding only the concepts with too few positive examples (fewer than 10) to adequately train the models. The resulting concept models are used as baseline concept detectors, which can be leveraged for context fusion in enhancing concept detection or search or applied directly in concept-based search (discussed further in Section 3).

3. MULTIMODAL SEARCH

In multimodal search, the objective is to retrieve a list of shots matching a semantic information need. The primary difference between multimodal search and concept detection is the amount of supervision. In search, only a text query and a few example images are given. In this work, we implement a baseline multimodal search system using no example images at all, a scenario which is highly desirable for the user and likely to be adopted. The system retrieves relevant video shots with queries based on text sentences or keywords and is built upon two primary search methods: text search and concept-based search. The two search methods are applied independently and their resulting scores are fused.

In **text search**, we issue simple keyword-based searches over documents composed of speech recognition transcripts associated with the videos. The videos are automatically segmented into semantic story units. Shots found to be temporally within the boundaries of a story are scored based on the textual similarity between the query and the story text.

In **concept-based search**, the text keywords are mapped to text keywords associated with the set of 374 pre-trained concept detectors. For each text keyword, the set of relevant concepts is found by searching lists of pre-defined keywords for each concept. A query for "boat" would match up with the "ship" concept, etc. Given the large set of concepts, it is likely that a single text keyword may match multiple concept detectors, some of which may be unrelated or weak. We therefore select only a single concept for each text keyword, choosing the concept which is most frequent and has the highest-performing concept detector. This approach to incorporating concept detection results is very conservative, using at most one concept detector per provided keyword.

Finally, **class-dependent fusion** is used to combine the text and concept-based search methods. The resulting scores are combined using a weighted average, where the weight of each method is dependent upon the *class* (or type) of the query. For example, queries in the "named persons" class rely entirely on the text search method, while queries in the "sports" class use both methods evenly. The multimodal search system is discussed in greater detail in [5].

4. CONTEXT FUSION AND RERANKING: OVERALL APPROACH

The primary focus of this work is to draw upon a large reserve of concept detectors, such as the pool of 374 detectors discussed in Section 2 to uncover contextual cues to refine the individual concept detectors themselves and to provide broader context for multimodal search. While in concept detection it may be feasible to find contextual relationships between concepts using a fully supervised approach over the large sets of training data that are available, this approach will not be applicable in search, where no training data is available. We therefore need to leverage the light search models and mine these for context.

We propose to accomplish this by using initial scores resulting from baseline concept detectors (for concept detection) or simple search methods (for multimodal search queries). Specifically, we assume that the scores from the initial method are reasonably usable, and then take the top-returned results to be pseudo-positives and sample pseudo-negatives from the lower-ranked results, thereby deriving a pseudo-labeling for the target semantic concept, T, which is simply binary: pseudo-positive or pseudo-negative. For these images, we also have detector scores, C, for each of the concepts in the lexicon. C quantizes the normalized scores of the concept detectors into 20 bins, which is empirically found to be a reasonable number. The objective, then, is to learn the best way to improve the performance of the initial search method by looking at the relationships between the target pseudo-labels, T, and the concept detector scores, C. The first step that we take is to find a subset of concepts in the lexicon which have strong relationships with the target pseudo labels by measuring the mutual information between the two:

$$I(T;C) = \sum_{T} \sum_{C} P(T,C) \log \frac{P(T,C)}{P(T)P(C)}, \qquad (1)$$

where $P(T,C)$, $P(T)$, and $P(C)$ are all estimated by counting frequencies in the sampled set of shots. Once we have this knowledge of T, C, and the shared mutual information, $I(T;C)$, between the two for every concept, we can then move on to leverage C to improve the accuracy of T. This can be done using any given approach, such as treating the concept scores, C, as term-frequencies, like in text, and fitting the information into a traditional PRFB framework. Or, we might attempt to feed the feature space into other reranking frameworks [10]. We opt to employ an SVM-based approach, since SVMs have been shown time and again to be effective in learning light-weight models for image retrieval [15] and context fusion [16].

We form a feature vector for each subshot consisting of the scores for the concepts found to be related to the target. This feature vector is used as an input space for an SVM model, using the target as the class labels. We observe that a space consisting of all 374 concepts is too large for learning models, so the space is cut to only contain the concepts with the highest mutual information with the target. Experiments on a validation set show that roughly 75 concepts gives reasonable performance, so we fix this parameter across all experiments. The exact set of 75 concepts is chosen independently for each query, however. Since TRECVID video search is evaluated over the top-1000 shots, we select the top 1200 subshots (which typically encompass the top-1000 shots) as the pseudo-positive set. The pseudo-negatives are chosen randomly from the non-pseudo-positve subshots to give 3600 pseudo-negative examples, (roughly 3x as many negatives so as not to incur problems due to unbalanced training data). The sizes of the pseudo-positive and psuedo-negative sets are chosen through some preliminary experiments on a validation set. Future work may investigate a more rigorous mechanism for choosing these examples. The set of examples is randomly divided into three folds and the SVM is learned on two of the folds and tested on the third. This process is repeated three times, with each fold being held out for testing once. Each of the initial 1200 pseudo-positive subshots is scored in this process. The resulting score is then averaged with the initial

Category	Base	BCRF	SVM	Rerank
Program (3)	.591	.591 *0.0%*	**.610 *3.3%***	.609 *3.1%*
Setting (13)	.438	.475 *8.3%*	.472 *7.8%*	**.482 *10%***
People (8)	.404	**.443 *9.7%***	.418 *3.7%*	.418 *3.5%*
Objects (8)	.280	.287 *2.4%*	.301 *7.2%*	**.302 *7.9%***
Activities (2)	.238	.269 *12%*	**.274 *14%***	.260 *8.8%*
Events (2)	.471	.471 *0.0%*	.495 *5.1%*	**.548 *16%***
Graphics (2)	**.378**	.378 *0.0%*	.363 *-4%*	.369 *-2%*
All (39)	.399	.422 *5.8%*	.421 *5.5%*	**.427 *7.0%***

Table 1: Comparison of context-fusion techniques across a validation set selected from the TRECVID 2005 development data. Mean average precisions for each technique are shown across each class of concept.

score and returned as the final reranking result.

The process is lightweight and highly general. We will see that it is actually comparable to supervised methods in the concept detection task and provides significant improvement in the TRECVID automatic video search task.

5. EXPERIMENTS

5.1 Data Set

We conduct our experiments using the data from the NIST TRECVID 2005 and 2006 video retrieval benchmarks [1], which includes over 300 hours of broadcast news video from English, Chinese, and Arabic sources. The data is accompanied with speech recognition and machine translation transcripts in English. In each year, 24 query topics are provided with ground truth relevance labels collected through a pooling process which are distributed after the benchmark.

Our experiments also rely on the LSCOM [2] and LSCOM-Lite [14] concept lexicons. The LSCOM concept lexicon is a set of 449 visual concepts which were annotated over an 80-hour subset of the TRECVID data. The LSCOM-Lite lexicon is an early version of LSCOM, which is essentially a subset of 39 concepts thought to be the most essential.

5.2 Reranking for Concept Fusion

As an initial test for the reranking method, we apply it, along with several supervised context fusion methods for the detection of the 39 LSCOM-Lite [14] concepts over a validation set sampled from the TRECVID 2005 development data. We study the performance of using context of 374 concept detectors to improve the detection accuracy of each of he 39 LSCOM-lite concepts. The supervised context fusion methods applied are the Boosted Conditional Random Field (BCRF) and the SVM method. The BCRF is exactly as described in Section 1.2.1 and reported in [11], while the SVM method is essentially an adaptation of the Discriminative Model Fusion (DMF) approach reported in [16], modified to deal with our large-scale concept lexicon. The SVM method treats the concept detection scores as an input feature space to an SVM, but also reduces the total number of features by selecting only high-information concepts, using mutual information as discussed in Section 4. The primary difference between the SVM method and the Reranking method is that the reranking method performs feature selection and learning using only pseudo-labels, while the SVM method requires ground-truth labels. All three methods improve upon the baseline method discussed in Section 2.

Table 1 shows the performance of the various context fusion methods broken down across various concept categories. The performance is expressed in terms of non-interpolated average precision, a popular and standard metric in information retrieval and concept detection which estimates the area under the precision-recall curve. We see that all three methods provide similar improvements over the baseline method. This is notable since the reranking method requires none of the supervision necessary for the BCRF and SVM methods. The primary reason for this in this case, is that training data is very limited and the majority of it is necessary for training the baseline models, leaving little room for discovering appropriate concept fusion approaches. The reranking method gives many noisy examples, while the supervised methods must rely on very sparse true examples. The effects of the noisiness or sparseness seem to offset each other.

Figure 2 shows the performance of the reranking approach and the baseline for each of the 39 concepts. We see a consistent, though small, improvement for nearly all concepts, indicating that the approach is stable and robust.

5.3 Reranking for Search

Having seen in the previous section that the reranking method is stable and comparable to supervised approaches, we can move on with confidence to the more interesting task of applying it to multimodal search, which requires a new method like reranking, since large amounts of training data are never available for search.

Figure 3 shows the results of applying the reranking method to text, concept-based, and fused searches for each query in the TRECVID 2005 and 2006 automatic search task. Again, in nearly every case, we see small but significant and steady increases in nearly every query topic, on average improving upon the baseline by between 15% and 30%. For the dominant group of search topics (Concept), the proposed method achieves an encouraging improvement of 12%-15%.

Varying levels of improvement can be influenced by a number of factors, such as the quality of the initial search results (results that are too noisy will not give enough information to meaningfully rerank, while extremely strong results will be difficult to improve upon) and the availability of untapped concept detectors to reinforce the search result. The effects of these factors are discussed in the following section.

5.3.1 Class-Dependency

The multimodal fused search result (and reranking) shown at the bottom of Figure 3 is generated using a different weighted summation of the text and concept-based search scores depending upon the class of the query. In this application, we use five pre-defined query classes: **Named Person**, **Sports**, **Concept**, **Person+Concept**, and **Other**. Examples from each class are given below. Each incoming query is automatically classified into one of these classes using some light language processing, like part-of-speech tagging, named entity extraction, or matching against keyword lists, as described in [5]. The performance of each search method over each class is shown in Table 2.

Named Person queries (such as "Find shots of Dick Cheney") frequently have the most room for improvement by the reranking method, especially when the initial text search results are strong (as is the case in the TRECVID 2005 set). These queries rely solely on text search arrive at an initial ranking. Luckily, though, text search tends to give a very strong initial result, with many positives appearing near the top of the list. These results are also loaded with false pos-

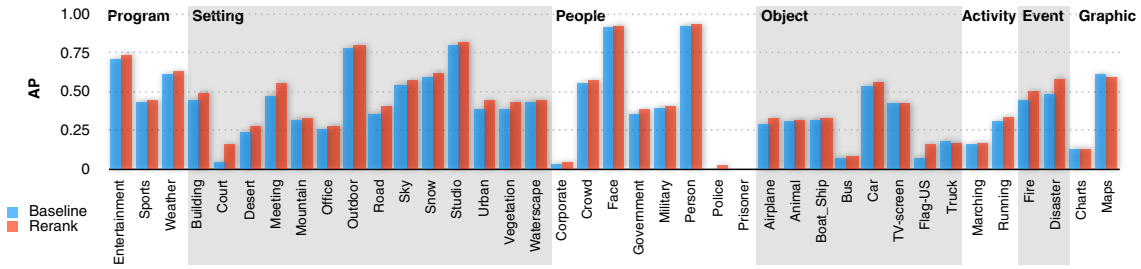

Figure 2: Average precision of baseline concept detectors and reranking concept detectors for the 39 LSCOM-lite concepts across a validation set selected from the TRECVID 2005 development data. Ordered by classes of concepts.

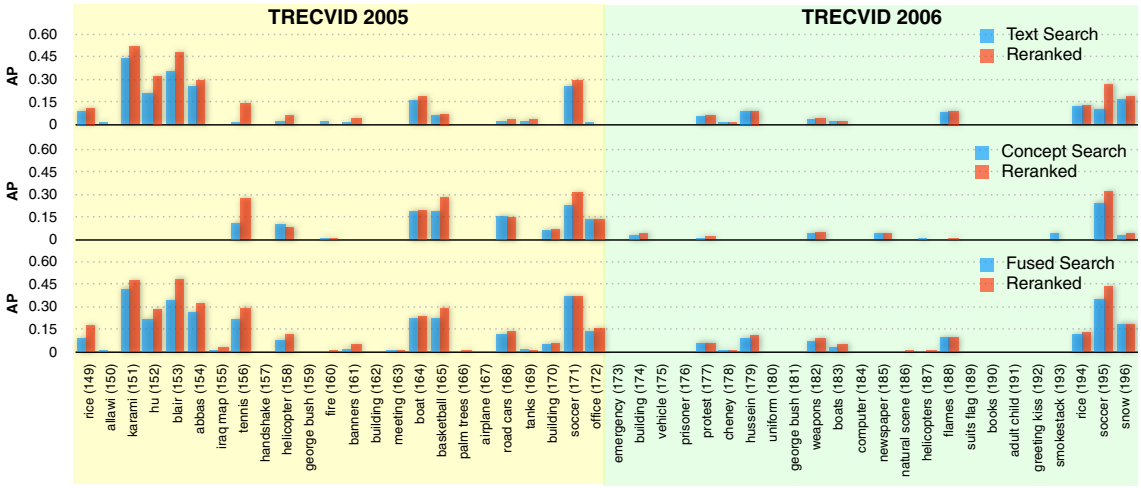

Figure 3: Average precisions of baseline and reranked search methods for each query in TRECVID 2005 and 2006. Text search shown in the top graph. Concept-based search in the middle. Fused multimodal search in the bottom.

itives, since all shots within a matching story are blindly returned. The application of pseudo-labeling can help sort this out, since shots relevant to the the query will likely come from the same news event featuring the sought-after person reported across various news sources. This will result in the occurrence of repeating scenes with shared contextual concept clues, while the false positives in the top-returned results will probably be indistinguishable from the pseudo-negatives that were sampled, resulting in the false positives being (correctly) pushed down the list.

Sports queries (like "Find shots of basketball courts") also have significant room for improvement from reranking. These queries rely fairly equally on concept-based and text search methods and the two are highly complimentary. When reranking is applied to sports queries, we expect a behavior much like the named person queries: the initial result is already very strong, so the context clues discovered via reranking are highly reliable. However, we also observe that the results from the initial search (particularly in the fused case) can be very strong, leaving little room (or need) for improvement via reranking.

Concept queries are quite different in character from either named person or sports queries. These queries are found to have keywords matching concepts in our lexicon (such as "Find shots of boats") and therefore rely mostly on concept-based search with some weight also on text search. We might expect that the degree of improvement for each concept-type query provided by reranking over the concept-based search, alone, might be similar to the improvement observed in the concept detection task as explored in Section 5.2; however, concept-type queries actually tend to be quite different from just the combination of one or two known concepts. They may also contain keywords which are not matched to any concepts, so it can be difficult to build a strong initial search result using concept-based search alone. In fact, we see that concept-based search for these queries is not improved significantly by reranking. However, if concept-based search is fused with text search, a stronger initial result is obtained for the query and reranking provides improvements over the fused multimodal result of about 15%.

Person+Concept queries (which includes queries matching the criteria for *both* the named person and concept classes, like "Find shots of George Bush leaving a vehicle") and **Other** queries (which includes any query not meeting the criteria for any of the other four classes) are harder to draw conclusions for, due to the limited number of queries fitting in to these classes. Given our examples, it seems that these two classes still lack viable methods for getting an initial search result, yielding a poor baseline and making it difficult to discover related concepts. These queries have low performance and reranking does not offer improvement.

5.3.2 *Feature Selection and Related Concepts*

Beyond simply observing the degrees in improvement experienced through reranking in context-based concept fusion, it is also interesting to examine the exact sources and manifestations of the contextual relationships between concepts and query topics. As mentioned in Equation 1, the

Class	Set	#	Text Search			Concept-based Search			Fused Multimodal		
			base	rerank	% imp.	base	rerank	% imp.	base	rerank	% imp.
Named Person	TV05	6	0.231	0.293	26.7%	0.000	0.000	0.0%	0.232	0.294	26.7%
	TV06	4	0.065	0.070	8.0%	0.000	0.000	0.0%	0.065	0.070	8.0%
Sports	TV05	3	0.116	0.174	50.0%	0.182	0.297	62.8%	0.276	0.325	17.8%
	TV06	1	0.109	0.268	145.5%	0.251	0.326	29.8%	0.358	0.445	24.3%
Concept	TV05	11	0.029	0.032	12.6%	0.066	0.066	0.0%	0.064	0.074	15.7%
	TV06	16	0.029	0.033	14.3%	0.019	0.019	1.1%	0.037	0.042	12.6%
Person + Concept	TV05	2	0.007	0.002	-68.5%	0.003	0.002	-19.3%	0.012	0.018	48.0%
	TV06	0	0.000	0.000	0.0%	0.000	0.000	0.0%	0.000	0.000	0.0%
Other	TV05	2	0.013	0.026	99.0%	0.003	0.004	25.6%	0.014	0.035	146.2%
	TV06	3	0.002	0.002	0.0%	0.015	0.015	0.0%	0.002	0.002	0.0%
All	TV05	24	0.087	0.112	28.7%	0.054	0.068	27.1%	0.124	0.153	22.9%
	TV06	24	0.033	0.042	27.4%	0.024	0.028	14.3%	0.049	0.056	15.0%

Table 2: Breakdown of improvements in mean average precision (MAP) over various query classes on the TRECVID 2005 (TV05) and TRECVID 2006 (TV06) search tasks. Shows the number of queries in each class (#), with the MAP of the baseline method (base), reranking result (reranking), and relative improvement (% imp.) over each of the available unimodal tools (text search and concept-based search) and the fused multimodal result.

Original Query	Positive Concepts	Negative Concepts
(151) Find shots of Omar Karami, the former prime minister of Lebannon.	Government Leader, Adult, Meeting, Furninture, Sitting	Overlaid Text, Commercial Advertisement, Female Person
(152) Find shots of Hu Jintao, president of the People's Republic of China.	Asian People, Government Leader, Suits, Group, Powerplants	Commercial Advertisement, Flags, Single Female Person, Logos Full Screen
(158) Find shots of a helicopter in flight.	Exploding Ordnance, Weapons, Explosion Fire, Airplane, Smoke, Sky	Person, Civilian Person, Face, Talking, Male Person, Sitting
(164) Find shots of boats or ships.	Waterscape, Daytime Outdoor, Sky, Explosion Fire, Exploding Ordnance	Person, Civilian Person, Face, Adult, Suits, Ties
(179) Find shots of Saddam Hussein.	Court, Politics, Lawyer, Sitting, Government Leader, Suits, Judge	Desert, Demonstration or Protest, Crowd, Mountain, Outdoor, Animal
(182) Find shots of soldiers or police with weapons and military vehicles.	Military, Machine Guns, Desert, Rocky Ground, Residential Buildings	Single Person, Corporate Leader, Actor, Female Person, Entertainment

Table 3: Concepts found to be strongly positively or negatively correlated to some example queries through reranking.

degree of correlation between a target semantic concept and any given concept detector in the lexicon is determined by measuring the mutual information between the two, giving concepts that are both negatively and positively correlated with the target concept. We can further distinguish between positively and negatively correlated concepts by utilizing the *pointwise mutual information*:

$$I_P(t;c) = \log \frac{P(t,c)}{P(t)P(c)}, \quad (2)$$

where if $I_P(t = \text{pseudo-postive}; c = \text{postive})$ is greater than $I_P(t = \text{pseudo-postive}; c = \text{negative})$, then the concept is considered to be positively correlated with the semantic target. Otherwise, it is considered to be negatively correlated. This approach has been used in prior works to determine the utility of concepts in answering search queries [13]; however that analysis was applied on ground truth relevance labels and concept annotations. In this work, we are measuring pseudo-labels and automatic concept detection scores.

Some interesting examples of concepts found to be related to query topics are shown in Table 3. Scanning through this table, we can observe concept relationships coming from a number of different mechanisms.

The first, and perhaps most obvious, type of relationship is essentially the discovery of **generally present** relationships, such as "Government Leader" for "Hu Jintao" and "Omar Karami" or "Waterscape" for "boats or ships." These are the relationships that we would expect to find, as many search topics have direct relationships with concepts and many topics might only occur in settings with a specific concept present. Virtually all of the negatively correlated concepts also fall into this class (or a variant of the class for generally *not* present concepts). In fact, we see that the negative relationships are dominated by detectable production artifacts, such as graphics or commercials which are rarely positively associated with typical search topics.

The second type of relationship is a **news story** relationship. In this relationship, scenes containing the target topic occur as part of some news story, where additional related concepts can be discovered that are unique to this news story, but not generally true across all time frames. The named person queries typically display these relationships. The "Hu Jintao" topic is found to be related to the "Powerplants" concept, not because Hu Jintao is typically found in a powerplant setting, but because the search set contained news stories about a visit to powerplants. Similarly, the "Saddam Hussein" topic is found to be related to the "Court," "Judge," and "Lawyer" concepts, though this relationship is only true during the time frame of the search set, during which Saddam Hussein was on trial. Also, the "Find shots of boats or ships" topic is found to be related to the "Explosion Fire" concept. There are no examples of boats with fires in the training data; however, the search set contains a news story about an occurrence of a boat fire. This second class of contextual relationship is uniquely discoverable by only the reranking method, since external training sets are likely to be constrained in time and from different time periods, making it impossible to predict new relationships arising in emerging news stories.

A third type of relationship is a **mistaken** relationship.

In this relationship, the system can discover context cues from an erroneous concept detector, which end up being beneficial despite the mistaken relationship. For example, it is found that the "Helicopter" topic is related to the "Airplane" concept; however, this relationship is actually false: these two concept typically do not occur in the same scene. The "Airplane" concept detector is not perfect and it turns out that the errors it makes tend to include helicopters since both have similar low-level appearances, so this mistaken relationship between concepts ends up being correct and beneficial with respect to the effects of imperfect detectors.

6. CONCLUSIONS AND FUTURE WORK

We have presented a new framework for incorporating contextual cues from large sets of pre-computed concept detectors for refining and reranking concept detection and search results. The proposed model differs from conventional contextual fusion models in that it requires no training data for discovering contextual relationships and instead mines an initial hypothesis ranking to find related concepts and reorder the original ranking. This unsupervised implementation allows the framework to be applied to search results, where training data for contextual relationships is typically not present, and where past approaches have been far more conservative, using only a few concept detectors per query. The reranking approach enables robust utilization of dozens of concept detectors for a single query to discover rich contextual relationships. To the best of our knowledge, this is the first work using a large pool of 374 concept detectors for unsupervised search reranking.

We find that the reranking method is comparable to supervised methods for concept detection tasks, improving 7% over baseline concept detection in MAP. The success of the approach is also observed in search tasks, where the method improves 15%-30% in MAP on the TRECVID 2005 and 2006 search tasks. The method is particularly successful for "Named Person," "Sports," and "Concept" queries, where the initial search result is reasonably strong. The method is shown to discover many of the direct concept relationships that would likely be discoverable using a supervised approach with training data, while also discovering relationships that are entirely present only in the search set, due to the temporal growth and decay of news cycles, and would therefore not be discoverable in a temporally separated training set. The method does not improve over queries where the initial search results perform poorly. Future research should focus on these difficult queries.

7. ACKNOWLEDGMENTS

This research was funded in part by the U.S. Government VACE program. The views and conclusions are those of the authors, not of the US Government or its agencies.

8. REFERENCES

[1] NIST TREC Video Retrieval Evaluation http://www-nlpir.nist.gov/projects/trecvid/.

[2] LSCOM Lexicon Definitions and Annotations Version 1.0, DTO Challenge Workshop on Large Scale Concept Ontology for Multimedia. Technical report, Columbia University, March 2006.

[3] M. Campbell, S. Ebadollahi, M. Naphade, A. P. Natsev, J. R. Smith, J. Tesic, L. Xie, and A. Haubold. IBM Research TRECVID-2006 Video Retrieval System. In *NIST TRECVID workshop*, Gaithersburg, MD, November 2006.

[4] J. Carbonell, Y. Yang, R. Frederking, R. Brown, Y. Geng, and D. Lee. Translingual information retrieval: A comparative evaluation. *Proceedings of the Fifteenth International Joint Conference on Artificial Intelligence*, pages 708–715, 1997.

[5] S.-F. Chang, W. Hsu, W. Jiang, L. Kennedy, D. Xu, A. Yanagawa, and E. Zavesky. Columbia University TRECVID-2006 Video Search and High-Level Feature Extraction. In *NIST TRECVID workshop*, Gaithersburg, MD, November 2006.

[6] S.-F. Chang, W. Hsu, L. Kennedy, L. Xie, A. Yanagawa, E. Zavesky, and D. Zhang. Columbia University TRECVID-2005 Video Search and High-Level Feature Extraction. In *NIST TRECVID workshop*, Gaithersburg, MD, November 2005.

[7] T.-S. Chua, S.-Y. Neo, K.-Y. Li, G. Wang, R. Shi, M. Zhao, and H. Xu. TRECVID 2004 search and feature extraction task by NUS PRIS. In *TRECVID 2004 Workshop*, 2004.

[8] A. Hauptmann, M.-Y. Chen, M. Christel, C. Huang, W.-H. Lin, T. Ng, N. Papernick, A. Velivelli, J. Yang, R. Yan, H. Yang, and H. D. Wactlar. Confounded expectations: Informedia at TRECVID 2004. In *TRECVID 2004 Workshop*, 2004.

[9] A. G. Hauptmann, M.-Y. Chen, M. Christel, W.-H. Lin, R. Yan, and J. Yang. Multi-Lingual Broadcast News Retrieval. In *NIST TRECVID workshop*, Gaithersburg, MD, November 2006.

[10] W. Hsu, L. Kennedy, and S.-F. Chang. Video search reranking via information bottleneck principle. In *ACM Multimedia*, Santa Babara, CA, USA, 2006.

[11] W. Jiang, S.-F. Chang, and A. 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] W. Jiang, S.-F. Chang, and A. C. Loui. Context-based Concept Fusion with Boosted Conditional Random Fields. In *IEEE ICASSP*, 2007.

[13] W. Lin and A. Hauptmann. Which Thousand Words are Worth a Picture? Experiments on Video Retrieval Using a Thousand Concepts. July 2006.

[14] M. Naphade, L. Kennedy, J. Kender, S. Chang, J. Smith, P. Over, and A. Hauptmann. LSCOM-lite: A Light Scale Concept Ontology for Multimedia Understanding for TRECVID 2005. Technical report, IBM Research Tech. Report, RC23612 (W0505-104), May, 2005.

[15] A. P. Natsev, M. R. Naphade, and J. Tesic. Learning the semantics of multimedia queries and concepts from a small number of examples. In *ACM Multimedia*, pages 598–607, Singapore, 2005.

[16] J. Smith, M. Naphade, and A. Natsev. Multimedia semantic indexing using model vectors. *Multimedia and Expo, 2003. ICME'03. Proceedings. 2003 International Conference on*, 2, 2003.

[17] C. G. Snoek, M. Worring, D. C.Koelma, and A. W. Smeulders. Learned lexicon-driven interactive video retrieval. In *CIVR*, 2006.

[18] C. G. M. Snoek, J. C. van Gemert, T. Gevers, B. Huurnink, D. C. Koelma, M. V. Liempt, O. D. Rooij, K. E. A. van de Sande, F. J. Seinstra, A. W. M. Smeulders, A. H. Thean, C. J. Veenman, and M. Worring. The MediaMill TRECVID 2006 Semantic Video Search Engine. In *NIST TRECVID workshop*, Gaithersburg, MD, November 2006.

[19] R. Yan. *Probabilistic Models for Combining Diverse Knowledge Sources in Multimedia Retrieval*. PhD thesis, Carnegie Mellon University, 2006.

[20] R. Yan, A. Hauptmann, and R. Jin. Multimedia search with pseudo-relevance feedback. *Intl Conf on Image and Video Retrieval*, pages 238–247, 2003.

[21] A. Yanagawa, S.-F. Chang, L. Kennedy, and W. Hsu. Brief descriptions of visual features for baseline trecvid concept detectors. Technical report, Columbia University, July 2006.