SEMANTIC MULTIMEDIA RETRIEVAL USING LEXICAL QUERY EXPANSION AND MODEL-BASED RERANKING

Alexander Haubold [‡]
Department of Computer Science
Columbia University, New York, NY 10027
ahaubold@cs.columbia.edu

Apostol (Paul) Natsev, Milind R. Naphade
IBM Thomas J. Watson Research Center
Hawthorne, NY 10532
{natsev,naphade}@us.ibm.com

ABSTRACT

We present methods for improving text search retrieval of visual multimedia content by applying a set of visual models of semantic concepts from a lexicon of concepts deemed relevant for the collection. Text search is performed via queries of words or fully qualified sentences, and results are returned in the form of ranked video clips. Our approach involves a query expansion stage, in which query terms are compared to the visual concepts for which we independently build classifier models. We leverage a synonym dictionary and WordNet similarities during expansion. Results over each query are aggregated across the expanded terms and ranked. We validate our approach on the TRECVID 2005 broadcast news data with 39 concepts specifically designed for this genre of video. We observe that concept models improve search results by nearly 50% after model-based re-ranking of text-only search. We also observe that purely model-based retrieval significantly outperforms text-based retrieval on non-named entity queries.

1. INTRODUCTION

Semantic search and retrieval of multimedia content is a challenging research field that has drawn significant attention of the multimedia research community. With the dramatic increase in video data available through different channels of dissemination, offline and online, methods of effective indexing and search of visual content are vital in unlocking the value of the content. Conventional text search over large databases is a well-understood problem with ubiquitous applications. However, search in non-textual content, such as image and video data, being a relatively new field, is not explored to the same degree. It is now apparent that merely applying conventional text search techniques to video will not work and need to be extended to include the semantics of the video content. The most substantial work in this field is presented in the TREC Video Retrieval Evaluation (TRECVID[1]) community, which focuses its efforts on evaluating video retrieval approaches by providing common video datasets and a standard set of queries.

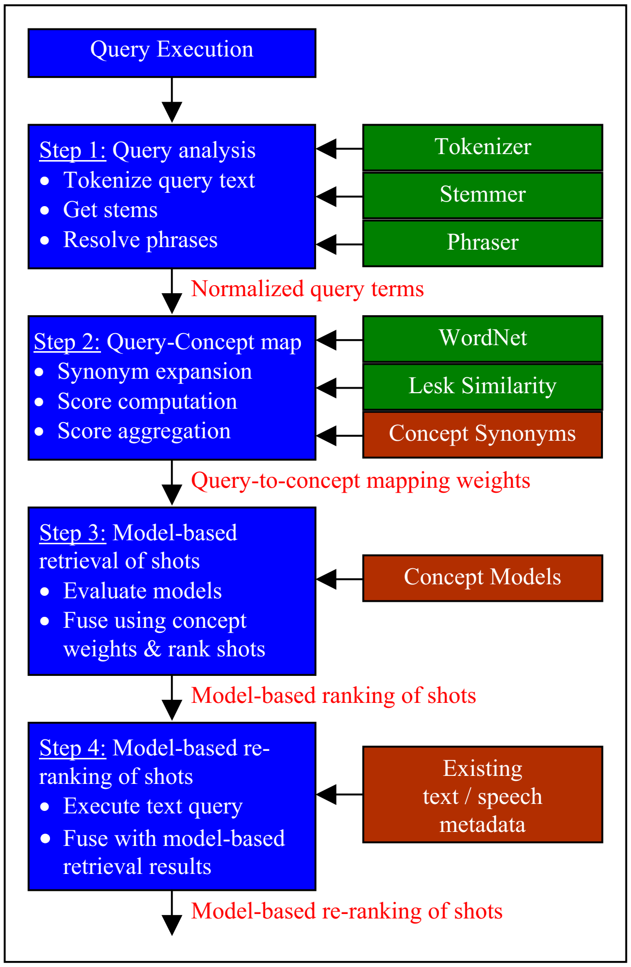
Query Execution
Step 1: Query analysis
• Tokenize query text
• Get stems
• Resolve phrases
Tokenizer
Stemmer
Phraser
Normalized query terms
Step 2: Query-Concept map
• Synonym expansion
• Score computation
• Score aggregation
WordNet
Lesk Similarity
Concept Synonyms
Query-to-concept mapping weights
Step 3: Model-based retrieval of shots
• Evaluate models
• Fuse using concept weights & rank shots
Concept Models
Model-based ranking of shots
Step 4: Model-based re-ranking of shots
• Execute text query
• Fuse with model-based retrieval results
Existing text / speech metadata
Model-based re-ranking of shots

Figure 1. Overview of model-based retrieval and re-ranking.

Our approach to multimedia search and retrieval addresses non-annotated broadcast news video data, for which speech transcripts may or may not be available. We assume that we have a set of models that can be applied to automatically detect a corresponding set of concepts such that each video shot can be annotated with a detection confidence score for each concept. Successful concept modeling and detection approaches have been developed in TRECVID, relying predominantly on visual analysis and statistical machine learning methods [1,2,6]. Our approach to search and retrieval leverages such concept models to

[‡] The work was done while the author was visiting the IBM T.J. Watson Research Center.

[1] http://www-nlpir.nist.gov/projects/trecvid

1-4244-0367-7/06/$20.00 ©2006 IEEE

enable or improve video search in scenarios with limited or no metadata. In particular, we focus on lexical query analysis and expansion—mapping query words and phrases to concepts—and we build a ranked list of matching shots based purely on automatic concept detection scores and automatically computed query-to-concept relevance scores (Figure 1). In addition, when textual metadata is available in the form of annotations, closed caption, automatic speech recognition, or video OCR transcripts, we use the model-based retrieval method to re-rank the purely text-based retrieval results. We validate our approach on the TRECVID 2005 corpus and query topics, and compare it to the text-based retrieval baseline. We observe 50% improvement in retrieval precision over the text-only baseline after model-based re-ranking. We also observe that for non-named entity queries, the model-based retrieval approach alone outperforms speech-based retrieval by 38%.

2. APPROACH

Our approach is split into four parts: 1. Query term extraction of words and qualified phrases, 2. Calculation of semantic relatedness scores between extracted query terms and concepts, 3. Model-based retrieval of shots through fusion across relevant concept detection results, and 4. Model-based re-ranking of shots through fusion of text-based and model-based retrieval results.

2.1. Semantic concept lexicon

For the multimedia research community, the TRECVID benchmark has succeeded in bringing semantic concept detection front and center. This has not only allowed different statistical learning techniques to be compared [6], but also sparked off a healthy debate on identifying a lexicon and a taxonomy that would be effective in covering a large number of queries. One such exercise to address the issue of a shallow taxonomy of generic concepts that can effectively address a large number of queries resulted in the creation of the LSCOM-lite lexicon (Figure 2).

As a first step we built support vector machine based semantic concept models [6] for all the annotated concepts of the LSCOM-lite lexicon based on some visual features from the training collection. Each of these models can then be used to get a quantitative score indicating the presence of the corresponding concept in any test set video shot. This quantitative score can then be converted into a confidence score which can in turn be used in our re-ranking experiments to modify the overall rank of a shot vis a vis its relevance to the query through the relationship of the concept to the query.

Since each concept can be described by multiple words and phrases with equivalent meaning, many of which may not be determined by WordNet as synonyms, we decided to manually create a synonym dictionary, which lists for each concept a number of similar words and phrases that represent that concept (Table 1).

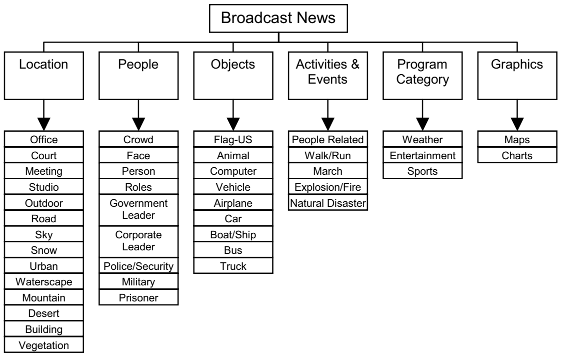

Figure 2. The LSCOM-lite Lexicon [7] designed for the TRECVID 2005 Benchmark consists of more than 40 concepts spread across multiple concept-types such as objects, events, sites etc. Of these, 39 concepts were annotated and made available for training models.

Concept	Representative words
airplane	air, aircraft, airline, carrier, helicopter, vehicle, warplane
natural disaster	disaster, earthquake, fire, flame, flood, hurricane, tornado, tsunami
sports	sport, baseball, basketball, soccer, tennis, cricket, football, hockey, golf, game, match
US flag	American flag, stars and stripes

Table 1. Sample concepts with synonyms. Synonyms are qualified WordNet terms, compared to query terms to resolve matching concepts.

2.2. Query analysis

We extract all individual words W from query Q, and add qualified WordNet phrases P from the same query:

$$\text{Extracted query terms } QT = \{W, P\}$$

A phrase is defined as the largest number of consecutive words that form a qualified phrase in WordNet. Phrases tend to disambiguate individual words and create a different, more specific and accurate meaning. We found it important to include such phrases in addition to words to form a more complete sense of the query. Examples for phrase extraction are presented in Table 2.

2.3. Query to concept mapping

In this step we compute each query term's concept weight vector, which is determined from a term's similarity to each concept. This vector is fused with shot concept confidences in the third step to produce a shot's relevance to the query. We use an *adapted Lesk semantic relatedness* score [3], one of many available for WordNet, to compute similarities between pairs of words. The adapted Lesk score between two terms is based on the amount of overlap (in words) between the definitions of the two terms, as well as between the definitions of their immediate neighbors according to WordNet relationships. Unlike other *semantic similarity* measures, the Lesk *semantic relatedness* measure considers not only *is-a* relationships between words (i.e., synonyms) but other relationships as well (e.g., *has-a*). This allows us to compute more general semantic relatedness scores between pairs of words or phrases, which are better suited

```
Let $S_1$ = senses($term_1$)
Let $S_2$ = senses($term_2$)
for all ($s_1$, $s_2$) where $s_1$ χ $S_1$, $s_2$ χ $S_2$
  $maxlesk$ = max(lesk($s_1$, $s_2$)) }
```

Algorithm 1. (*maxlesk*) Approach to sense disambiguation in computing the *Lesk* semantic relatedness score for two terms.

```
QT  = set of extracted query terms for query Q      Input
C   = set of concepts from fixed lexicon            Input
CS  = set of synonyms S for all concepts in C       Input
QC  = set of concept confidences for query Q        Output
for each $QT_i$ in QT
  for each $C_j$ in C
    for each $C_jS_k$ in $C_jS$
      $conf_k$ = maxlesk($C_jS_k$, $QT_i$) }        Step 1
    $QC_j$ += max($conf_k$) } }                     Step 2
for each $C_j$ in C
  $QC_j$ /= |QT| }                                  Step 3
```

Algorithm 2. Approach to computing weighted mapping from a given query to a set of semantic concepts.

```
S   =  set of shots                          Input
SC  =  set of concepts C for each shot S     Input
SS  =  shot score                            Output
for each $S_i$ in S
  for each $C_j$ in C
    $S_iS$ += $QC_j$ * $SC_j$ } }
```

Algorithm 3. Approach to computing the final ranked list of shots, given a query-to-concept mapping (with weights) and concept detection confidences for all shots.

taking off	Depart from the ground
military vehicle	Vehicle used by the armed forces
shaking hands	Take someone's hands and shake …
basketball player	An athlete who plays basketball

Table 2. Sample phrases in which individual words have vague or different meaning. *Phrasing* resolves more specific senses, which result in higher precision *Lesk* similarities due to better matching definitions.

Query: "*people with banners or signs*"	
T1: *people*	*people-marching*: 4737, *crowd*: 4737, …
T2: *banner*	*people-marching*: 216, *US flag*: 151, …
T3: *sign*	*building*: 5361, *waterscape/front*: 531, …
Final query to model mapping with weights: *Building*: 1879.7, *People/marching*: 1685.7, C*rowd*: 1625.3, …	

Table 3. Example of a complete query expansion with *Lesk* semantic relatedness scores.

for query expansion purposes. For example, the term *airplane* is semantically associated with the term *aircraft* through an *is-a* relationship but is also associated to other terms, such as *airline*, *pilot, to fly,* through different relationships. All of these terms are in fact suitable for query expansion purposes, even though they are not synonyms to the original query term, and may even be designated as different parts of speech (i.e., nouns, verbs, adjectives, etc.). Due to the above features, we use the adapted Lesk score to measure general semantic relatedness between a pair of terms. Since terms, however, may belong to multiple parts of speech (POS), and may carry multiple meanings, or senses, we perform a sense disambiguation step, which attempts to resolve the correct meanings for a pair of terms. In particular, we select the highest Lesk similarity score between their respective sets of senses (Algorithm 1) based on the intuition that the most similar senses are most likely to be the ones used in the same context. For an overview of other semantic similarity and semantic relatedness measures, see [4].

Query terms are compared to all synonyms of a concept (Algorithm 2, Step 1). For each (query term, concept) combination, we then record the highest similarity score (that of the best matching synonym) as the concept weight for the given query term, and aggregate these scores over all query terms (Algorithm 2, Step 2). We normalize the final query-to-concept weight scores by the number of query terms (Algorithm 2, Step 3), resulting in a weight vector that ranks all 39 concepts with respect to their semantic relatedness to the query. Of these, we currently consider only the top 3 most similar concepts (and their weights) for each query, and set all other weights to 0 to reduce noise effects. In the future, we plan to allow a variable number of concepts per query based on the query properties.

2.4. Model-based retrieval

In this step, query concept vectors (concept weight scores) are fused with shot concept vectors (concept detection results) to determine a ranked list of matching shots. For each shot's concept detection vector, we multiply the query concept weight vector, and use the sum of products as a final model-based retrieval score for the given shot and the given query (Algorithm 3). An example of a complete query expansion following the algorithm is presented in Table 3.

2.5 Model-based re-ranking

In the final (and optional) step, the model-based shot ranking is used to re-rank the results of text-based retrieval, if available. The intuition is that specific queries such as named entities cannot be answered precisely through generic models only (unless named entities are modeled explicitly), although the latter can serve to refine—and hopefully improve—retrieval results based on other modalities, such as closed caption (CC) or automatic speech recognition (ASR) transcripts, video OCR, etc. If such text sources are available, we therefore execute a text search against these sources, and we then fuse the results with the model-based retrieval results in order to get a re-ranked and possibly improved set of results. Ideally, we can use different fusion weights for the two approaches, depending on query topic characteristics, such as whether the topic is about a named entity or not. Such query-dependent fusion approaches [5] are very promising but require a separate training set of sample topics to determine the optimal fusion parameters. For simplicity, we consider only simple non-weighted score averaging (after global range normalization) as the preferred fusion method.

3. PERFORMANCE EVALUATION

We have evaluated the proposed approach on the TRECVID 2005 test corpus and query topics[2]. This collection contains 140 broadcast news video clips from U.S., Arabic, and Chinese sources, with durations of 30 minutes to 1 hour each, pre-segmented into 45,765 shots. Each video comes with a speech transcript obtained through automatic speech recognition, as well as machine translation for the non-English sources. The text search baseline is obtained with *JuruXML*—a text search engine available as part of the IBM UIMA SDK[3]. We use Average Precision at depth of 1000 to measure performance on a specific topic, and Mean Average Precision (MAP) to aggregate performance results across multiple topics. Average Precision is the official performance metric adopted by TRECVID, and essentially represents the area under the precision-recall curve.

The results on the 24 search topics are presented in Figure 3, which lists performance scores across all topics, denoted by representative query phrases. The TRECVID 2005 topics include 7 specific/named topics and 17 generic ones (unnamed objects, scenes, and events). Relative performance on these two classes differs significantly, as summarized in Table 4. From the results, it is evident that model-based retrieval improves substantially upon the text search baseline for generic topics but understandably fails at named entities. However, when fused with the text search baseline, model-based retrieval can effectively filter and re-rank shots, leading to consistent and substantial improvements for both query classes. Performance on named entity topics is improved by over 23%, while that on generic topics is improved by an incredible 89%! Overall, model-based re-ranking improves performance across all topics by nearly 50%, which is a clear testament to the promise of the proposed approach.

4. CONCLUSION AND ACKNOWLEDGEMENT

We have presented and evaluated a new approach to retrieval of visual information by leveraging visual models, applying query expansion, and re-ranking results in a fusion step. In query expansion, we compare visual models and their pre-defined synonyms to query terms using semantic relatedness. The resulting list of visual models is used in a fusion step with text-based retrieval methods to formulate a final ranked list of search results. Our evaluations show that this approach significantly improves retrieval for generic concepts, and after fusion with text-based retrieval improves retrieval over all topics, including named entities.

This material is based upon work funded in part by the U.S. Government. Opinions, findings and conclusions or recommendations expressed here are those of the authors and do not necessarily reflect the views of the U.S. Government.

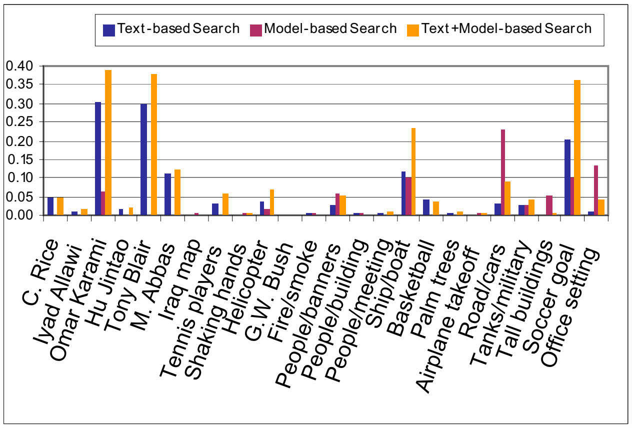

Figure 3. Performance evaluation (y-axis = Average Precision at 1000) of text-based search, model-based search, and model-based re-ranking on 24 TRECVID 2005 topics. Model-based re-ranking performed by simple averaging fusion of text-based and model-based retrieval results after score range normalization.

Query Class Specificity (topic count)	Text-Search Baseline	Model-Based Retrieval	Model-Based Re-ranking (gain)
Named (7)	0.113	0.010	0.139 (*23%*)
Unnamed (17)	0.032	0.044	0.061 (*89%*)
All Topics (24)	0.056	0.034	0.083 (*49%*)

Table 4. Performance summary (Mean Average Precision scores) for text-based retrieval baseline vs. proposed model-based retrieval and model-based re-ranking approaches. Performance scores aggregated over two complementary query classes—specific and generic topics, as well as over all topics.

5. REFERENCES

[1] A. Amir, J. Argillander, M. Campbell, A. Haubold, G. Iyengar, S. Ebadollahi, F. Kang, M.R. Naphade, A. Natsev, J.R. Smith, J. Tešić, T. Volkmer, "IBM Research TRECVID-2005 Video Retrieval System," *TRECVID Workshop*, Gaithersburg, MD, Nov, 2005.

[2] A. Natsev, M.R. Naphade, J. Tešić, "Learning the Semantics of Multimedia Queries and Concepts from a Small Number of Examples," *ACM Multimedia*, Singapore, pp. 598-607, Nov, 2005.

[3] S. Banerjee, T. Pedersen, "Extended Gloss Overlaps as a Measure of Semantic Relatedness," *Joint Conference on Artificial Intelligence*, Morgan K., Mexico, pp. 805-810, Aug. 9-15, 2003.

[4] S. Patwardhan, S. Banerjee, T. Pedersen, "Using Measures of Semantic Relatedness for Word Sense Disambiguation," *Conference on Intelligent Text Processing and Computational Linguistics*, Springer, Mexico, pp. 241-257, Feb. 16-22, 2003.

[5] L.S. Kennedy, A. Natsev, S.F. Chang, "Automatic Discovery of Query-Class-Dependent Models for Multimodal Search," *ACM Multimedia*, ACM Press, Singapore, pp. 882-891, Nov. 6-11, 2005.

[6] M. Naphade, J.R. Smith, F. Souvannavong, "On the Detection of Semantic Concepts at TRECVID," *ACM Multimedia*, ACM Press, New York, NY, pp. 660-667, Oct. 10-16, 2004.

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

[2] http://www-nlpir.nist.gov/projects/tv2005/tv2005.html

[3] http://www.alphaworks.ibm.com/tech/uima