Semantic Concept-Based Query Expansion and Re-ranking for Multimedia Retrieval*

A Comparative Review and New Approaches

Apostol (Paul) Natsev
IBM Thomas J. Watson
Research Center
natsev@us.ibm.com

Alexander Haubold
Dept. of Computer Science
Columbia University
ahaubold@cs.columbia.edu

Jelena Tesic
IBM Thomas J. Watson
Research Center
jtesic@us.ibm.com

Lexing Xie
IBM Thomas J. Watson
Research Center
xlx@us.ibm.com

Rong Yan
IBM Thomas J. Watson
Research Center
yanr@us.ibm.com

ABSTRACT

We study the problem of semantic concept-based query expansion and re-ranking for multimedia retrieval. In particular, we explore the utility of a fixed lexicon of visual semantic concepts for automatic multimedia retrieval and re-ranking purposes. In this paper, we propose several new approaches for query expansion, in which textual keywords, visual examples, or initial retrieval results are analyzed to identify the most relevant visual concepts for the given query. These concepts are then used to generate additional query results and/or to re-rank an existing set of results. We develop both lexical and statistical approaches for text query expansion, as well as content-based approaches for visual query expansion. In addition, we study several other recently proposed methods for concept-based query expansion. In total, we compare 7 different approaches for expanding queries with visual semantic concepts. They are evaluated using a large video corpus and 39 concept detectors from the TRECVID-2006 video retrieval benchmark. We observe consistent improvement over the baselines for all 7 approaches, leading to an overall performance gain of 77% relative to a text retrieval baseline, and a 31% improvement relative to a state-of-the-art multimodal retrieval baseline.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: Content Analysis and Indexing; H.3 [**Information Storage and Retrieval**]: Information Search and Retrieval; I.2 [**Artificial Intelligence**]: Learning—*Concept Learning*

*This material is based upon work supported by the US Government. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the US Government.

MM'07, September 23–28, 2007, Augsburg, Bavaria, Germany.
Copyright 2007 ACM 978-1-59593-701-8/07/0009 ...$5.00.

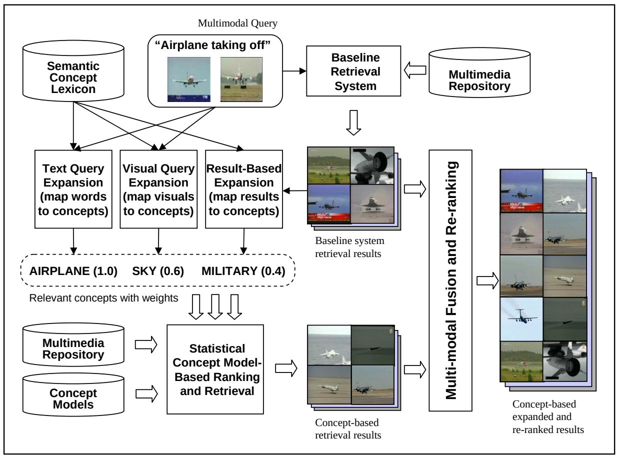
Multimodal Query
"Airplane taking off"
Semantic Concept Lexicon
Baseline Retrieval System
Multimedia Repository
Text Query Expansion (map words to concepts)
Visual Query Expansion (map visuals to concepts)
Result-Based Expansion (map results to concepts)
Baseline system retrieval results
AIRPLANE (1.0)
SKY (0.6)
MILITARY (0.4)
Relevant concepts with weights
Multimedia Repository
Concept Models
Statistical Concept Model-Based Ranking and Retrieval
Concept-based retrieval results
Multi-modal Fusion and Re-ranking
Concept-based expanded and re-ranked results

Figure 1: Overview of concept-based retrieval and re-ranking framework. Three general approaches are illustrated for identifying relevant semantic concepts to a query—based on textual query analysis, visual content-based query modeling, and pseudo-relevance feedback. A multi-modal fusion step leverages the relevant concepts to improve the results.

1. INTRODUCTION

Search and retrieval are vital parts of multimedia content management, and are increasingly receiving attention with the growing use of multimedia libraries and the explosion of digital media on the Web. By its virtue, multimedia spans multiple modalities, including audio, video, and text. While search and retrieval in the text domain are fairly well-understood problems and have a wide range of effective solutions, other modalities have not been explored to the same degree. Most large-scale multimedia search systems typically rely on text-based search over media metadata such as surrounding html text, anchor text, titles and abstracts. This approach, however, fails when there is no such metadata (e.g., home photos and videos), when the rich link structure of the Web cannot be exploited (e.g., enterprise content and archives), or when the metadata cannot precisely capture the true multimedia content.

On the other extreme, there has been a substantial body of work in the research community on content-based retrieval methods by leveraging the query-by-example and relevance feedback paradigms. These methods do not require any additional metadata, but rely on users to express their queries in terms of query examples with low-level features such as colors, textures and shapes. Finding appropriate examples, however, is not easy, and it is still quite challenging to capture the user intent with just a few examples.

To address these issues, a new promising direction has emerged in recent years, namely, using machine learning techniques to explicitly model the audio, video, and image semantics. The basic idea is that statistical detectors can be learned to recognize semantic entities of interest—such as people, scenes, events, and objects—by using more training examples than a user would typically provide in an interactive session (in most cases using hundreds or thousands of training examples). Once pre-trained, these detectors could then be used to tag and index multimedia content semantically in a fully automated fashion. Work in this field related to video concept detection and retrieval has been driven primarily by the TREC Video Retrieval Evaluation (TRECVID) community [37], which provides a common testbed for evaluating approaches by standardizing datasets, benchmarked concepts and queries [20, 1, 14, 35, 7]. A major open problem is how to scale this approach to thousands of reliable concept detectors, each of which may require thousands of training examples in turn. Notable efforts in creating large training corpora include the collaborative annotation efforts undertaken by TRECVID participants, donated annotations by the MediaMill team from the University of Amsterdam [34], as well as the Large Scale Concept Ontology for Modeling (LSCOM) effort to define and annotate on the order of 1000 concepts from the broadcast news domain [21]. At present, however, reliable concept detectors are still limited to tens of concepts only, while usable concept detectors exist for a few hundred concepts at best. It is therefore imperative to develop techniques that maximally leverage the limited number of available concept detectors in order to enable or improve video search when the metadata is limited or completely absent.

In this paper, we study the problem of semantic concept-based query expansion and re-ranking for multimedia retrieval purposes. In particular, we consider a fixed lexicon of semantic concepts with corresponding visual detectors, and we explore their utility for automatic multimedia retrieval and re-ranking purposes. We propose several new approaches for query expansion, in which the query textual terms, query visual examples, or the query baseline results are analyzed to identify relevant visual concepts, along with corresponding weights. The most salient concepts are then used to generate additional query results (improve recall) or to re-rank an existing set of results (improve precision).

Specific contributions. We propose a novel lexical query expansion approach leveraging a manually constructed rule-based mapping between a lexicon of semantic text annotations and a lexicon of semantic visual concepts. The proposed approach uses deep parsing and semantic tagging of queries based on question answering technology. It outperforms two popular lexical approaches based on synonym expansion and WordNet similarity measures. We also propose a novel statistical corpus analysis approach, which identifies significant correlations between words in the English language and concepts from the visual concept vocabulary based on their co-occurrence frequency in the video corpus. This approach performs on par with, or better than, lexical query expansion approaches but has the advantage of not requiring any dictionaries, or manually constructed concept descriptions, and is therefore more scalable. We also study smart data sampling and content modeling techniques for concept-based expansion based on visual query examples. Finally, we study and evaluate several other previously proposed methods for concept-based query expansion and retrieval. In total, we compare and empirically evaluate 7 different approaches for expanding *ad-hoc* user queries with relevant visual concepts. We observe consistent improvement over the baselines for all 7 approaches, leading to an overall performance gain of 77% relative to a text retrieval baseline, and a 31% improvement relative to a multimodal retrieval baseline. To the best of our knowledge, this is one of the most comprehensive reviews and evaluations of concept-based retrieval and re-ranking methods so far, and it clearly establishes the value of semantic concept detectors for answering and expanding ad-hoc user queries.

2. RELATED WORK

2.1 Text-Based Query Expansion

Since concept-based query expansion is related to research in text-based query expansion, we give an overview of the main approaches from that domain. In principle, the idea is to expand the original query with additional query terms that are related to the query. The addition of related terms can improve recall—especially for short queries—by discovering related documents through matches to the added terms. It may also refine the meaning of overly broad queries, thereby re-ranking results and improving precision. This of course works only as long as the refined query is indeed consistent with the original one. Experiments in text document retrieval have shown that query expansion is highly topic-dependent, however.

2.1.1 Lexical approaches (language-specific)

Lexical approaches leverage global language properties, such as synonyms and other linguistic word relationships (e.g., hypernyms). These approaches are typically based on dictionaries or other similar knowledge representation sources such as WordNet [38]. Lexical query expansion approaches can be effective in improving recall but word sense ambiguity can frequently lead to topic drift, where the semantics of the query changes as additional terms are added.

2.1.2 Statistical approaches (corpus-specific)

Statistical approaches are data-driven and attempt to discover significant word relationships based on term co-occurrence analysis and feature selection. These relationships are more general and may not have linguistic interpretation. Early corpus analysis methods grouped words together based on their co-occurrence patterns within documents [28]. Related methods include term clustering [17] and Latent Semantic Indexing [10], which group related terms into clusters or hidden orthogonal dimensions based on term-document co-occurrence. Later approaches attempt to reduce topic drift by looking for frequently co-occurring patterns only within the same context, as opposed to the entire document, where the context can be the same paragraph, sentence, or simply a neighborhood of n words [32, 16, 40, 12, 5].

2.1.3 Statistical approaches (query-specific)

In contrast to global statistical approaches, which consider the distribution and co-occurrence of words within an entire corpus, *local analysis* uses only a subset of the documents to identify significant co-occurrence patterns. This subset is typically a set of documents explicitly provided or tagged by the user as being relevant to the query. In *relevance feedback* systems, for example, the system modifies the query based on users' relevance judgments of the retrieved documents [31]. To eliminate or reduce the need for user feedback, some systems simply assume that the top N retrieved documents are relevant, where N is determined empirically and is typically between 20 and 100. This is motivated by the assumption that the top results are more relevant than a random subset and any significant co-occurrence patterns found within this set are more likely to be relevant to the query. This approach is called *pseudo-relevance feedback*, or *blind feedback* [40].

2.2 Visual Concept-Based Query Expansion

Existing approaches for visual concept-based retrieval can similarly be categorized into the three categories for text-based approaches—lexical and global statistical or local statistical approaches. There are a few differences, however, in that the documents in this case are multimodal (e.g., video clips) as opposed to purely textual, and the correlations of interest involve visual features or concepts, as opposed to just words. Also, multimodal queries have additional aspects than text queries, and can include content-based query examples. Existing approaches can therefore also be broadly categorized depending on the required input—query text, visual query examples, or baseline retrieval results. Table 1 summarizes related work along both dimensions.

2.2.1 Lexical approaches (language-specific)

Lexical approaches for visual concept-based query expansion are based on textual descriptions of the visual concepts, which essentially reduce the problem to that of lexical text-based query expansion. Typically, each concept is represented with a brief description or a set of representative terms (e.g., synonyms). Given a textual query, the query words are then compared against the concept descriptions, and any matched concepts are used for refinement, where the matching may be exact or approximate [7, 35, 13, 24, 8]. In the latter case, lexical similarity is computed between the query and the concepts using WordNet-based similarity measures, such as Resnik [30] or Lesk [2, 25]. Alternatively, Snoek et al. [35] also consider the vector-space model for similarity between queries and concept descriptions. In [35], word sense disambiguation is performed by taking the most common meaning of a word, as the authors found this to be the best approach from a number of disambiguation strategies they considered. In [13], the authors disambiguate term pairs by taking the term senses that maximize their pairwise Lesk similarity. In [8], the term similarities are modeled as a function of time to reflect the time-sensitive nature of broadcast news and to filter out "stale" correlations.

2.2.2 Statistical approaches (corpus-specific)

To the best of our knowledge, there is no previous work on global corpus analysis approaches for visual concept-based query expansion. We propose one such method in Section 4.3.

Concept-based query expansion approaches	Text query expansion	Visual query expansion	Result-based expansion
Lexical (language-specific)	[7, 35, 13, 8] Sec. 4.1–4.2		
Statistical (corpus-specific)	Section 4.3		
Statistical (query-specific)		[33, 22, 35, 29, 36] Sec. 4.4–4.5	[42] Sec. 4.6

Table 1: Summary of concept-based query expansion and retrieval approaches categorized along two dimensions. See text for description of approaches.

2.2.3 Statistical approaches (query-specific)

Local statistical approaches for visual concept-based query expansion have not been explored as much as lexical approaches. Several works [33, 22, 29, 36] consider content-based retrieval where the query is specified with one or more visual examples represented by semantic feature vectors. They project each example onto a semantic vector space where the dimensions represent concept detection confidences for images/shots. Once queries are mapped to the semantic space, traditional content-based retrieval techniques are used. We previously proposed the probabilistic local context analysis method described in Section 4.6 [42].

3. CONCEPT-BASED RETRIEVAL AND RE-RANKING FRAMEWORK

The proposed concept-based retrieval and re-ranking framework is illustrated in Figure 1. We consider approaches for concept-based expansion of text queries, visual content-based queries, as well as initial retrieval results produced by any baseline retrieval system, be it text-based, content-based, or multimodal. The identified relevant concepts are then used to retrieve matching shots based on statistical detectors for the 39 LSCOM-lite concepts used in TRECVID.

3.1 Semantic concept lexicon

For the multimedia research community, TRECVID provides a benchmark for comparison of different statistical learning techniques. It 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) used in the TRECVID Video Retrieval evaluation [37]. Recently, the full LSCOM effort has produced a concept lexicon of over 1000 visual concepts for the broadcast news domain, along with annotations for many of them over a large video corpus [21]. This annotated corpus contains nearly 700K annotations, over a vocabulary of 449 concepts and a video set of 62K shots. Independently, the MediaMill team from the University of Amsterdam has donated a lexicon and annotations for 101 semantic concepts [34] defined on the same corpus. These large, standardized, and annotated corpora are extremely useful for training large sets of visual concept detectors and allow better comparability across systems.For the experiments in this paper, we use the 39 LSCOM-lite concepts from Figure 2 since they have the most robust detectors and are the ones used in TRECVID 2005–2006.

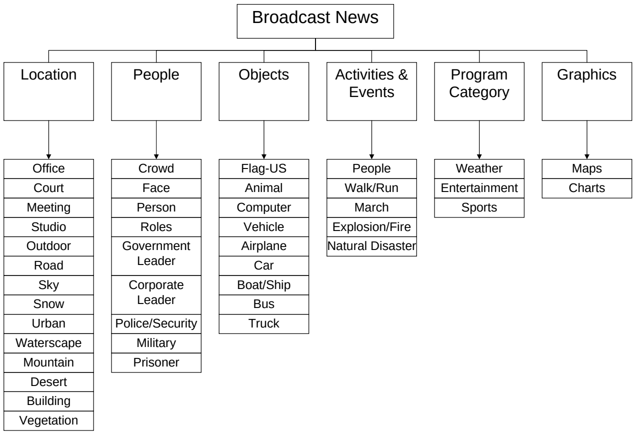

Figure 2: The LSCOM-lite concept lexicon.

As a first step we built support vector machine models for all 39 concepts of the LSCOM-lite lexicon based on low-level visual features from the training collection [1, 4]. These models are used to generate quantitative scores indicating the presence of the corresponding concept in any test set video shot. Quantitative scores are converted into confidence scores, which are used in our re-ranking experiments. The resulting concept detectors achieve some of the best scores in the TRECVID High Level Feature Detection task [1, 4].

Since each concept encompasses broad meaning and can be described by multiple words and phrases, we manually describe each concept with a set of synonyms and other words and phrases that represent its meaning. We also manually map these representative concept terms to WordNet synsets to remove any ambiguity in their meaning. This is a commonly used approach in the literature [7, 1, 13, 8, 35]. Each concept in our lexicon therefore has a brief text description, a set of WordNet synsets, and a corresponding detector.

3.2 Query analysis

For textual queries, we perform common query analysis and normalization, including stop-word removal and Porter stemming. We also identify phrases using a dictionary lookup method based on WordNet, as well as statistical part-of-speech tagging and sense disambiguation based on deep parsing and context modeling. We further filter queries to keep only nouns, verbs, and phrases. In addition, named entities are extracted and query terms are annotated using a semantic text annotation engine, which detects over 100 named and nominal semantic categories, such as named/nominal people, places, organizations, events, etc. [3].

For visual content-based queries, we extract low-level visual descriptors from all query examples, and evaluate each concept model with respect to these descriptors. This results in a 39-dimensional model vector of concept detection scores for each visual example.

3.3 Concept-based query expansion

After query analysis, we apply one of 7 concept-based query expansion methods to identify a set of relevant concepts for the query, along with their weights. These methods are described in more detail in Section 4. Some of these methods have been previously proposed (Sections 4.1, 4.4, and 4.6), while others are new (Sections 4.2, 4.3, and 4.5).

3.4 Concept-based retrieval

In this step, we use the identified concepts, along with their relevance weights, to retrieve relevant results from the video corpus. In particular, each concept detector is used to rank the video shots in order of detection confidence with respect to the given concept. Given the set of related concepts for a query, the corresponding concept detection ranked lists are combined into a single concept-based retrieval result list. We use simple weighted averaging of the confidence scores, where the weights are proportional to the query-concept relevance score returned for each concept.

Concept confidence normalization. Approaches for confidence score normalization based on voting schemes, range, or rank normalization are frequently used in meta-search or multi-source combination scenarios [9, 19]. Other approaches designed specifically for normalizing concept detection scores use score smoothing based on concept frequency or detector reliability [33, 35, 15]. Finally, approaches for calibrating raw SVM scores into probabilities are also relevant but typically require additional training data [27]. For our experiments we use statistical normalization of the confidence scores to zero mean and unit standard deviation:

$$S'_c(x) = \frac{S_c(x) - \mu_c}{\sigma_c}, \qquad (1)$$

where $S_c(x)$ is the confidence score for concept c on shot x, and μ_c and σ_c are the collection mean and standard deviation of the confidence scores for concept c.

3.5 Concept-based expansion/re-ranking

In this step, the concept-based retrieval results are used to expand/re-rank the results of the baseline retrieval system. We differentiate between two fusion scenarios—expanding/re-ranking the results of a baseline retrieval system *vs.* fusing the results of multiple retrieval systems from multiple modalities for the final result ranking. To reduce complexity and alleviate the need for a large number of training queries needed to learn fusion parameters, we use simple non-weighted averaging in the case of baseline result expansion and re-ranking. This approach typically generalizes well and has proven to be robust enough for re-ranking purposes.

3.6 Multi-modal/multi-expert fusion

Finally, we fuse the results across all available modalities and retrieval experts using query-dependent fusion techniques. This is motivated by the fact that each retrieval expert typically has different strengths/weaknesses. For example, text-based retrieval systems are good at finding named entities, while content-based retrieval systems are good for visually homogeneous query classes, such as sports, weather. We consider linear combination for multiple rank-lists, and use fusion weights that depend on the properties of the query topic. Such query-dependent fusion approaches have been found to consistently outperform query-independent fusion [41, 18], and we use a novel variant of query mapping that would generate query-classes dynamically.

We use a semantic text annotation engine [3] to tag the query text with more than one hundred semantic tags in a broad ontology, designed for question-answering applications on intelligence and news domains. The set of over one hundred tags covers general categories, such as person, geographic entities, objects, actions, events. For instance, ”Hu Jintao, president of the People’s Republic of China” is tagged with "Named-Person, President, Geo-Political Entity, Nation". We manually design a mapping from all semantic tags to seven binary feature dimensions, intuitively described as *Sports, Named-Person, Unnamed-Person, Vehicle, Event, Scene, Others.* This mapping consists of rules based either on commonsense ontological relationships, e.g., "President" leads to *Named-Person*, or on frequent concept co-occurrence, such as "Road" implies *Vehicle*.

The query text analysis and feature mapping is performed on both the set of known (training) queries as well as on each of the new queries. We map a new query to a small set of *neighbors* among the training queries (with inner product of query features serving as the similarity measure), and then dynamically generate optimal query weights for those training queries on-the-fly. This dynamic query fusion scheme performs better than hard or soft query-class-dependent fusion in our case, since the query feature space is relatively clean and rather low-dimensional [39].

4. QUERY EXPANSION APPROACHES

In the following, we describe several concept-based query expansion approaches, which are evaluated in Section 5.

4.1 Lexical Concept Matching

The first two approaches we consider are the naive word-spotting approach, or matching query terms to concept descriptions directly, as well as a lexical query expansion approach based on a WordNet similarity measure between visual concept descriptions and text queries. These are by far the most commonly used approaches for visual concept-based query expansion [1, 7, 35, 13, 4, 6, 8]. With the naive approach, using manually established concept descriptions consisting of representative concept "trigger" words, we find direct matches between query terms and visual concepts (e.g. term "flight" resolves to concept *Airplane*, "soccer" resolves to *Sports*, etc.). While only applicable to a limited number of queries, this approach produces a highly accurate mapping between visual and textual modalities.

In a separate and a more generally applicable query expansion approach, the related concepts to a query are identified not only by exact matches between query terms and concept descriptions, but also by soft matching and similarity scoring based on WordNet relatedness between visual concept descriptions and query terms. In particular, we follow the approach from [13], where visual concepts are weighted based on their similarity to query terms through an adapted Lesk semantic relatedness score [2, 25]. The Lesk score is computed for all pairs of query terms and concept representative terms. For a given term-concept pair, we select the highest Lesk similarity score based on the intuition that the most similar senses are most likely to be the ones used in the same context. Similarity score vectors of 39 concepts for each query term are aggregated and normalized by the number of query terms, resulting in the query's lexical similarity to the 39 visual concepts. In a departure from [13], we have a final de-noising step, where we threshold all similarity scores (i.e., concept weights), and keep only the significant weights that are larger than the mean plus 1 standard deviation, as calculated over the 39-dimensional concept weight vector. We find that this de-noising step generalizes better than the global top-k weight thresholding mechanism used in [13], as it results in a variable number of related concepts per query, which is more realistic for general queries.

4.2 Lexical Rule-based Ontology Mapping

Complimentary to the lexical word-based concept matching described in the previous section, we also utilize deep semantic analysis of the query text and derive a mapping from the semantic tags on the words/phrases to visual concepts. Deeper text analysis is beneficial, since words can have multiple senses, and direct word-spotting may not distinguish them. For instance, in the query *people in uniform and in formation*, the words "uniform" and "formation" have very distinct meanings, which are not apparent without considering the part-of-speech information and the context. In this case, a specially designed mapping from text semantics to visual concepts is useful, since (1) some semantic text categories have a closely-related visual category (e.g., *Named-Person → Person, Face*), while others may not have direct visual implications (e.g., *Nation*); (2) even when there is a strong text-to-visual connection, the mapping is rarely one-to-one, due to the currently limited ontology for both text and visual categories. For example, text annotations *Vehicle, Road* shall map to *Car, Bus*, while *Vehicle, Waterbody* shall map to *Boat_Ship* in the LSCOM ontology.

Similar to Section 3.6, we use a semantic text annotation engine [3] to tag the query text with more than one hundred semantic tags in a broad ontology. A set of a few dozen rules are manually designed to map the semantic tags to one or more of the LSCOM-lite concepts, each of which belongs to one of the following three types (1) one-to-one or one-to-many mappings, e.g., a text tag "Sport" maps to visual concepts *Sports* and *Walking_Running*; (2) many-to-one or many-to-many mappings, e.g., "Vehicle" and "Flight" imply *Sky, Airplane*; and (3) negated relationships, such as "Furniture" implies *NOT Outdoors.* We note that even though this mapping is manual, it is not query dependent since it maps one fixed ontology (of text annotations) to another fixed ontology (of visual annotations). When both ontologies are in the order of hundreds of concepts, this is feasible to do manually, and provides a higher quality mapping than automatic mapping approaches based on WordNet for example, as seen in Section 5.1). However, when either ontology grows to thousands of concepts, this approach becomes less feasible. In such cases, we expect that statistical automatic mapping approaches, such as the one described in Section 4.3, will be the most feasible ones. In fact, as shown in Section 5.4, these approaches can outperform the lexical mapping approaches, even if the latter use a manually constructed ontology mapping.

4.3 Statistical Corpus Analysis

Global corpus analysis approaches for query expansion typically perform correlation analysis between pairs of terms based on their co-occurrence counts in the corpus. Term co-occurrence can be measured within the same document, the same paragraph/sentence, or within a small window of a few neighboring words only. Given the term co-occurrence counts, a statistical test is usually performed to measure which correlations are significant. Identified term pairs with significant correlations are then linked together so that when either term appears in the query, the other can be used for query expansion purposes.

We adopt the same methodology for visual concept-based query expansion, except that we identify significant correlations between words in the English language vocabulary and visual concepts in the semantic concept lexicon (LSCOM-lite). To measure co-occurrence counts, we implicitly associate the visual concepts detected for a given video shot with all of the words from the video speech transcript that occur within a fixed temporal neighborhood around the given shot. To this end, we applied a likelihood ratio statistical significance test called the $\mathcal{G}^2$ test [11]. Dunning [11] showed that this test is more accurate than Pearson's χ^2 test, especially for sparse contingency tables, and introduced the test to the computational linguistics community where it is now widely used. It can be shown that $\mathcal{G}^2(X, Y)$ can also be expressed in terms of *mutual information* $I(X; Y)$ as follows:

$$\mathcal{G}^2(X, Y) = 2N \cdot I(X; Y) = 2N\left(H(X) + H(Y) - H(X, Y)\right),$$

where $H(\cdot)$ is the entropy function. Mutual information is considered highly effective for feature selection purposes [43], and the above formula shows that the $\mathcal{G}^2$ statistic produces a proportional value, with the advantage that it can be evaluated for statistical significance using a χ^2 distribution table with 1 degree of freedom.

For the experiments in this paper, we use the $\mathcal{G}^2$ test to identify strong correlations between terms from the speech transcript and visual concepts. All $\mathcal{G}^2$ scores are thresholded using a certain significance level (e.g., 95%, 99%, 99.9%, or 99.99% confidence interval), and the remaining scores are normalized into Cramer's ϕ correlation coefficients:

$$\phi(X, Y) = \sqrt{\mathcal{G}^2/N}. \qquad (2)$$

Using the above correlations, we build an associative weighted map between speech terms and visual concepts. Given an arbitrary query, all concepts that are strongly correlated to any of the query terms are then used for query expansion purposes using weights proportional to the corresponding ϕ correlation coefficients. In Section 5.2 we show cross-validation results for various confidence intervals used for $\mathcal{G}^2$ score thresholding, various concept binarization schemes, as well as a final performance comparison with the baseline.

4.4 Statistical Content-Based Modeling

This approach formulates the topic answering problem as a discriminant modeling one in a concept model vector space. The concept model vector space is constructed of concept detection confidences for each shot [33, 22]. If the query is specified with one or more visual examples, we map each visual example onto the concept model vector space, and approach the problem as content-based retrieval. Unlike low-level descriptor spaces, the concept model vector space is highly non-linear, however, due to the use of different modeling approaches and parameters, and is also non-orthogonal, due to correlations among concepts (e.g. *Sky* and *Mountain*). Standard content-based retrieval approaches based on simple nearest neighbor modeling do notwork very well in this space. Instead, we use Support Vector Machine (SVM) modeling with nonlinear kernels in order to learn nonlinear decision boundaries in this highly skewed space. Due to the high dimension of the space the limited number of distinct positive examples, we adopt a pseudo-bagging approach. In particular, we build multiple primitive SVM classifiers whereby the positive examples are used commonly across all classifiers but each has a different sampled set of pseudo-negative data points. The SVM scores corresponding to all primitive SVM models are then fused using Boolean AND logic to obtain a final model. Figure 3 illustrates the main idea. For more details, see [23, 36].

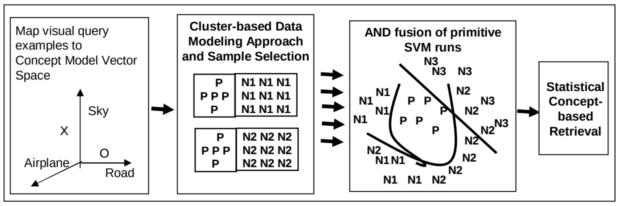
Map visual query examples to Concept Model Vector Space
Sky
X
O
Airplane
Road
Cluster-based Data Modeling Approach and Sample Selection
P
P P P
P
N1 N1 N1
N2 N2 N2
AND fusion of primitive SVM runs
N1
N2
N3
P
Statistical Concept-based Retrieval

Figure 3: Statistical concept-based modeling.

4.5 Statistical Content-Based Model Selection

Content-Based Model selection uses a set of statistical hypothesis tests to determine if a specific concept detector is relevant to the query based on the visual query examples. This is similar to the statistical corpus analysis-based approach from Section 4.3 but in this case we are looking for significant correlations between the set of visual query examples for a given topic and the concepts, as opposed to correlations between query terms and concepts. We use standard statistical hypothesis tests, such as t-score, χ^2 test, majority vote, one sided student test, and likelihood ratio statistical test to evaluate the importance and uniqueness of each concept with respect to the query examples. For one sided tests, we use the concept score distribution over the test set as a background distribution. For two-side tests, we use the concept score distribution over all query topic examples in the database as the hypothesis to test against. Thus, given the subset of related concepts for a specific query, and their weights, the corresponding concept detection ranked lists are combined using weighted averaging of the confidence scores.

4.6 Statistical Pseudo-Relevance Feedback

Pseudo-relevance feedback (PRF) examines initial retrieved documents per query topic as pseudo-positive/negative examples in order to select expanded discriminative concepts to improve the retrieval performance. Current PRF approaches use a small number of top-ranked documents to feed an automatic feedback process. If the initial retrieval performance returns poor results, as common in multimodal search scenarios, PRF is very likely to degrade the retrieval results. We take a more robust feedback approach, termed probabilistic local context analysis(pLCA) [42], that automatically leverages useful high-level semantic concepts to improve the initial retrieval output without assuming the top ranked examples are mostly relevant. The pLCA approach is derived from the probabilistic ranking principle, and it suggests ranking the retrieved shots in a descending order of the conditional probability of relevance. As combination weights for the semantic concepts are unknown, we treat them as latent variables w_c and incorporate them into a relevance-based probabilistic retrieval model. If M_T is the number of top-ranked shots defined by users, and $S_0(x_j)$ is the initial retrieval score for the shot x_j, we can compute the conditional probability of relevance y by marginalizing the latent variables w_c.

$$p(\mathbf{y}|S_0) \propto \int_{w_c} \prod_{j=1}^{M_T} \exp\left(y_j S_0(x_j) + y_j \sum_c w_c S_c(x_j)\right) dw_c, \quad (3)$$

We adopt the mean field approximation [26] to compute this marginal probability. First, we construct the family of variational distributions, $q(\mathbf{y}, w_c) = \prod_j q(w_c|\beta_c) \prod_j q(y_j|\gamma_j)$, as a surrogate to approximate the posterior distribution

$p(\mathbf{y}|S_0)$, where $q(\nu_c|\beta_c)$ is a Gaussian distribution with mean β_c and the variance $\sigma = 1$, and $q(y_j|\gamma_j)$ is a Bernoulli distribution with a sample probability of γ_j. After some derivations [42], we can find that the variational distribution closest to $p(\mathbf{y}|S_0)$ must satisfy the following fix point equations,

$$\begin{aligned} \gamma_j &= \left[1 + \exp\left(2S_0(x_j) + 2\sum_c \beta_c S_c(x_j)\right)\right]^{-1} \\ \beta_c &= \sum_j (2\gamma_j - 1) S_c(x_j). \end{aligned} \tag{4}$$

These equations are invoked iteratively until the change of KL-divergence is small enough. Upon convergence (which is almost always guaranteed), we use the final $q(y_j|\gamma_j)$ as a surrogate to approximate the posterior probability without explicitly computing the integral, and we simply rank the documents in a descending order of the parameter γ_j as the retrieval outputs. This iterative update process typically converges in a small number of iterations and thus it can be implemented efficiently in a real retrieval system.

5. COMPARATIVE EVALUATION

We evaluate all approaches described in the previous section on the TRECVID 2005 and 2006 test corpora and query topics [37]. Specifically, we use the TRECVID'05 corpus for parameter tuning and cross-validation, as needed, and evaluate final performance on TRECVID'06 corpus. Both collections consist of broadcast news video from U.S., Arabic, and Chinese sources, with durations of 30 minutes to 1 hour each. The TRECVID'05 test set consists of approximately 80 hours of video segmented into 45,765 shots, while the TRECVID'06 corpus has approximately 160 hours of video, segmented into 79484 shots. Each video comes with a speech transcript obtained through automatic speech recognition (ASR), as well as machine translation (MT) for the non-English sources. The quality of the ASR and MT transcripts is generally not very reliable but it is representative of the state of art in these fields and it is quite helpful for retrieval purposes, especially for named entity queries. For baseline comparison purposes, we use the speech-based retrieval approach described in [4]. Each data set comes with ground truth for 24 query topics. We use Average Precision at depth 1000 to measure performance on a specific topic, and Mean Average Precision (MAP) to aggregate performance across multiple topics. Average Precision is the performance metric adopted by TRECVID, and essentially represents the area under the precision-recall curve. Three example query topics and corresponding concepts identified by the approaches we consider are listed in Table 4.

5.1 Experiment I: Lexical query expansion

In the first experiment we evaluate the performance of the proposed lexical query expansion method leveraging a rule-based ontology mapping between a text annotation lexicon and the visual LSCOM-lite concept lexicon. We compare this approach against the text retrieval baseline as well as the two other lexical approaches for query expansion—synonym-based expansion (exact match) and WordNet-based expansion (soft match with Lesk-based similarity).

The results on both TRECVID collections are presented in Table 2, which lists performance of the text-based retrieval baseline and the three lexical query expansion approaches for three query classes and across all topics.

Query Class (# of topics)	Text- Only	Synonym match	WordNet similarity	Ontology Mapping
TRECVID-2005 Mean Average Precision				
PersonX (7)	0.217	0.228	0.204	**0.244**
People (9)	0.066	**0.1037**	0.092	0.088
Scenes (8)	0.036	0.061	**0.068**	0.067
All Topics (24)	0.099	0.124	0.116	**0.125**
TRECVID-2006 Mean Average Precision				
PersonX (4)	**0.148**	0.127	0.123	0.133
People (12)	0.018	0.032	**0.036**	0.028
Scenes (8)	0.055	0.062	0.058	**0.073**
All Topics (24)	0.052	0.058	0.058	**0.060**

Table 2: Performance summary (Mean Average Precision scores) for text-based retrieval baseline and three lexical query expansion approaches.

From the results, it is evident that all of the lexical concept-based query expansion approaches improve upon the text search baseline, and the proposed ontology rule-based mapping approach achieves the most significant improvements of the three lexical approaches. Improvement on the 2005 corpus ranges from a low of 12% on named person topics to a high of 85% on scene topics, with an average of about 27% gain over all topics. On 2006 data, the improvement is more modest, and we even see a loss of 10% on named entities. However, we can still observe substantial gains on the other query classes, including a 100% gain on generic people events, and a 34% gain on scene topics, leading to an overall improvement of 17% across all topics. From the experiments we also note that the three lexical approaches are somewhat complementary since top performance on each query class is typically achieved by different approaches. The proposed ontology mapping-based approach is most consistent, however, and achieves the best overall performance on both the 2005 and 2006 datasets.

5.2 Experiment II: Statistical corpus analysis

We have evaluated the proposed statistical corpus analysis-based method for text query expansion using 2005 and 2006 TRECVID data sets and query topics. First, we empirically determine parameter values such as the thresholding method for concept confidence binarization, and the $\mathcal{G}^2$-score confidence-level thresholding for concept selection per query. Visual concept detection is inherently probabilistic and therefore its values are continuous. For statistical mapping of concepts to related text we binarize these values using a threshold as a function of the collection mean, μ, and standard deviation, σ, of concept confidence scores (see Section 3.4). In the end, we select the mean + 2 standard deviations as a threshold but we observe that the method is quite robust with respect to other parameter settings. The other parameter we consider in our experiments determines the significance threshold for computed $\mathcal{G}^2$ correlation scores between a visual concept and a text term. We tried various significance levels based on the χ^2 distribution with 1 degree of freedom, and finally select a confidence interval of 99.9%. Again, we note that the method is robust with respect to this parameter, and the optimal setting generalizes on both collections (data not shown due to space considerations).

The final performance results are summarized across 3

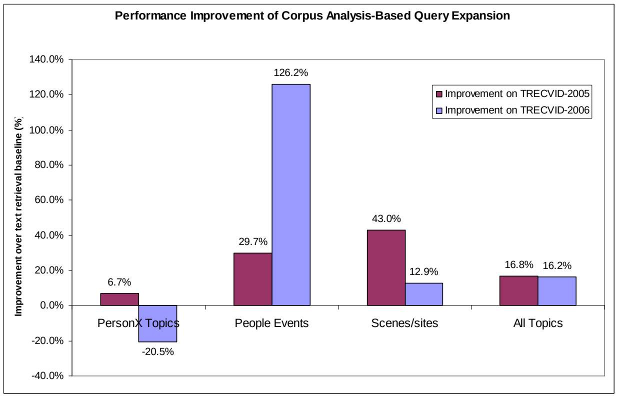

Figure 4: Performance improvement of statistical corpus analysis-based query expansion relative to text retrieval baseline on TRECVID 2005 and 2006.

query classes and over all topics in Figure 4. On average, concept-based expansion improves the baseline retrieval result by 16% and 17% for the TRECVID 2005 and 2006 collections respectively. We note that specific person queries benefit the least and in fact, deteriorate on the 2006 set, but the other query classes improve significantly, with 30%–40% gains on 2005 and over 100% gains on 2006 non-named people event queries. The limited improvement (or even loss) on person-X topics can be explained by the fact that the named entity queries are very specific so the potential contribution of generic visual concepts is limited. The performance loss on the TRECVID06 queries is most likely due to the fact that the 3 named entity queries in TRECVID 2006 were simply requests for a specific person, without additional constraints, such as entering/leaving a vehicle, a building, etc. In contrast, some of the named entity queries in 2005 included additional visual constraints, which benefited from visual concept-based filtering. Unlike the named entity query class, the general “people events” category benefited significantly from concept-based expansion and filtering on the 2006 data. The significantly larger performance gain, as compared to the same category on TRECVID 2005 data, is most likely due to the fact that the 2006 topics and data set were generally more difficult than their 2005 counterparts, which lowered the performance of the speech-based retrieval baseline and increased the dependency on other modalities and retrieval methods, such as concept-based retrieval and re-ranking. In general, the empirical results confirm that concept-based query expansion and retrieval is very topic-dependent. In Section 5.4, we show however that the query-specific nature of the improvement can be leveraged by our query-dependent fusion approach, which can adaptively select and fuse the best query expansion approaches for each topic, leading to further performance improvements.

5.3 Experiment III: Content-based Modeling

In this experiment, we evaluate the overall impact of content-based semantic query expansion and re-ranking as compared to a multimodal baseline *MM-baseline*. The *MM-baseline* is formed using query-independent fusion on speech-based and visual content-based retrieval runs to create a joint text-visual baseline (for details of both retrieval approaches, see [4]). The mean average precisions for TRECVID 2005 and 2006 datasets are shown in Table 3. The content-based semantic query modeling component, termed *MM-content* run is constructed as a fusion of the multimodal baseline and the approach described in Sec. 4.4. The improvement of the *MM-content* run over the baseline is significant over a range of topics, and results in 25% overall MAP gain for the 2006 dataset. The content-based semantic model selection component, termed *MM-selection* run is constructed as a fusion of the multimodal baseline and the approach from Sec. 4.5. We evaluated T-score, chi^2, majority vote, one-sided student test, and likelihood ratio statistical tests for LSCOM-lite concept relevance to the query topics from the TRECVID 2005 dataset. Thresholds were fixed on the global level based on the 95% statistical confidence level. T-score had the highest MAP, and the fusion with the multimodal baseline offered 6.7% improvement over the baseline. This corresponding improvement on the 2006 query topics was 8.2%, as shown in Table 3. The most relevant concept identified for the sample TRECVID 2006 topics using the T-score method are presented in Table 4.

TRECVID	MM-baseline	MM-content	MM-selection
2005	0.15143	0.16678	0.16169
2006	0.06962	**0.08704**	0.07535

Table 3: Content-based modeling impact on mean average precision (MAP) over the optimized multimodal baseline for TRECVID 2005 and 2006 topics for content-based semantic query modeling approach (Sec. 4.4) and content-based semantic model selection approach (Sec. 4.5).

5.4 Experiment IV: Overall Comparison

Table 4 lists three sample query topics from TRECVID’06 and their corresponding concepts identified by the approaches we considered. It can be observed that although these methods found some common concepts with each other, a number of unique concepts are identified by various approaches. For example, text-based approaches tend to identify semantically related concepts for the query, and their relation is easy to be interpreted by human. On the other hand, statistical-based approaches and visual-based approaches can find other visually related concepts such as *US_Flag, TV_Screen* for the ”Dick Cheney” query, *Vegetation* for the “soccer” query, which are difficult to discover by textual relations alone. However, due to the noisy learning process, it is also possible for the last three approaches to introduce unexpected noise in the results, such as *Weather* for the ”Dick Cheney” query in the pLCA method.

In order to quantitatively analyze the effects of model-based retrieval, we evaluate the overall system performance for all the model-based retrieval experts in two stages, where the first stage only considers the textual query, and the second stage takes multimodal queries including image examples into account. Figure 5 (left) compares the MAP of the individual and fused retrieval experts. *T-Baseline* is the text retrieval system based on speech transcripts only. *T-pLCA* re-ranks this baseline using probabilistic local context analysis and the LSCOM models as described in Section 4.6. *T-Lexical* and *T-Corpus* uses globally weighted linear combination of the two sets of model-scores (Sections 4.2 and 4.3)

Query Topic	Lexical WordNet Lesk similarity	Lexical rule-based ontology mapping	Statistical global corpus analysis	Prob. local context analysis	Visual content selection
U.S. Vice President Dick Cheney	Gov._Leader, Corp._Leader	Gov. Leader, Face	Corp._Leader, Studio, Gov._Leader, Meeting, PC_TV_Screen, Face	Corp._Leader, Weather Police, Vegetation PC_TV_Screen, US_Flag	Corp._Leader, Face, Gov._Leader, Person, Studio, US_Flag
Soldiers, police, guards escort a prisoner	Military, Court, Prisoner, Police, Person	Military, Court, Prisoner, Police, Person, Explosion	Military, Corp. Leader, US_Flag, Studio, Gov. Leader, PC_TV_Screen	Court	Military, Police, Prisoner, Gov. Leader, Crowd, Walking_Running
Soccer goalposts	Sports, Walking_Running	Sports, Person, Walking_Running	Sports, Vegetation, Walking_Running	Sports, Vegetation Walking_Running	Sports, Vegetation, Walking_Running

Table 4: Sample topics and most related concepts identified by various concept-based expansion approaches.

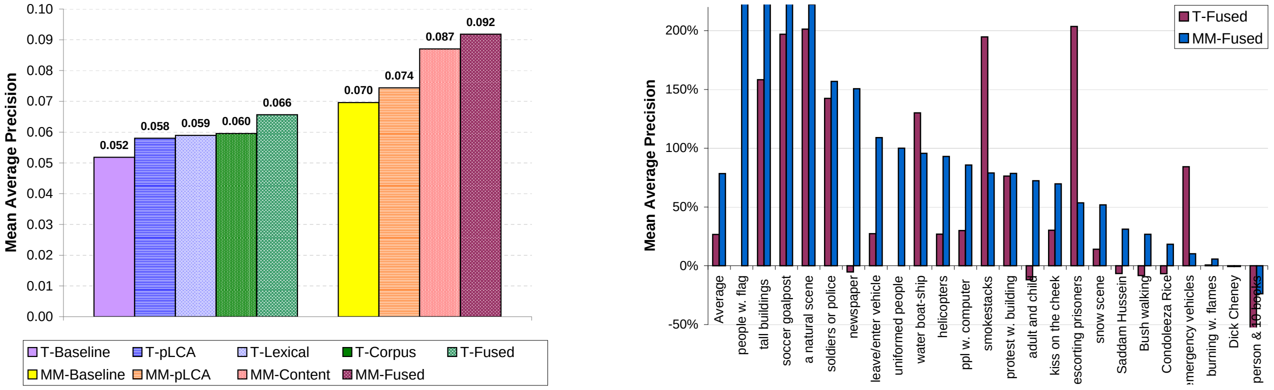

Figure 5: Performance evaluation of multimodal fusion on text-based retrieval systems and multi-modal retrieval systems. (left) MAP on TRECVID06 dataset and query topics, see Section 5.4 for definitions of the runs. (right) Relative improvements over text retrieval baseline for each of the 24 TRECVID06 queries.

and the baseline text runs, respectively. We can see that each of pLCA, Lexical and Statistical approaches improves the text retrieval baseline by 12% ∼ 15% (over a MAP of 0.052). Even though the 3 approaches perform similarly to each other, they are complementary, as query-dependent fusion of the four model-based retrieval runs brings the total improvement to 26%. Also, the proposed global corpus analysis-based approach performs on par with, or better than, all other approaches for text query expansion, without requiring any manual creation of synonym lists or mapping rules for each concept, as needed for the lexical approaches. This makes it the most scalable and generally applicable approach for concept-based text query expansion, and we expect it to significantly outperform lexical approaches on general topics which are not so easily linked to the concepts via synonyms (the TRECVID topics are somewhat biased towards the LSCOM-lite concepts by design).

In the second stage, we use query-independent fusion on text and content-based visual retrieval runs to create a joint text-visual baseline (*MM-Baseline*, which is 34% better than the text baseline. pLCA is again used to re-rank the baseline results to generate *MM-pLCA*. *MM-Content* augments *MM-baseline* by taking into account the concept presence of the visual examples as described in Section 4.4, resulting in 25% improvement over baseline alone for 2006 dataset. *MM-fused* is produced by query-dependent fusion on the three MM-runs, generating 0.092 in MAP, or equivalently, a 77% improvement over the text-only baseline and a 31% gain over the multimodal baseline. In both experiments, pLCA can consistently bring a small improvement over the baseline approaches, but this improvement is not as significant as the *MM-Content* run in terms of average precision.

Figure 5 (right) shows the relative improvement in average precision of the *T-fused* and *MM-fused* runs for each of the 24 TRECVID06 queries sorted by the resulting improvement in *MM-fused*. We can see that the queries that map well to existing LSCOM concepts with high-performing detectors (e.g., *Sports* concept for query *soccer goalpost*) are indeed the ones with the most improvement, while the lack of related semantic detectors hampers the improvement on some other queries, e.g., the last query *people with 10 books*. These observations clearly confirm the advantage of leveraging additional semantic concepts over the text/image retrieval. Moreover, the complementary information provided by various expansion methods also allows us to further improve the retrieval results by combining their outputs.

6. CONCLUSION

In this paper we considered the problem of semantic concept-based query expansion leveraging a set of visual concept detectors for video retrieval purposes. We presented a comprehensive review of existing approaches and proposed two novel methods based on lexical rule-based ontology mapping and statistical global corpus analysis. The statistical method is the first global corpus analysis-based method proposed for visual concept-based query expansion purposes, and performs on par with, or better than, the popular lexical approaches, without requiring manually constructed concept descriptions or mapping rules. All approaches were evaluated on TRECVID datasets and query topics, and resulted in significant improvements over state-of-art retrieval baselines that do not use concept models. In particular, we observed 77% improvement over a text-based retrieval baseline and 31% improvement over a multimodal baseline.

7. 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. Tevsić, and T. Volkmer. IBM Research TRECVID-2005 video retrieval system. In *NIST TRECVID Video Retrieval Workshop*, Gaithersburg, MD, Nov. 2005.

[2] S. Banerjee and T. Pedersen. Extended gloss overlaps as a measure of semantic relatedness. In *Joint Conference on Artificial Intelligence*, pages 805–810, Mexico, Aug. 9–15 2003.

[3] J. C.-Carroll, K. Czuba, J. Prager, A. Ittycheriah, and S. B.-Goldensohn. IBM's PIQUANT II in TREC2004. In *NIST Text Retrieval Conference (TREC)*, pages 184–191, Gaithersburgh, MD, USA, 16–19 November 2004.

[4] M. Campbell, A. Haubold, S. Ebadollahi, M. Naphade, A. Natsev, J. R. Smith, J. Tevsić, and L. Xie. IBM Research TRECVID-2006 Video Retrieval System. In *NIST TRECVID Video Retrieval Workshop*, Gaithersburg, MD, Nov. 2006.

[5] D. Carmel, E. Farchi, Y. Petruschka, and A. Soffer. Automatic query refinement using lexical affinities with maximal information gain. In *25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 283–290, Tampere, Finland, Aug. 11–15 2002.

[6] 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 Video Retrieval Workshop*, Gaithersburg, MD, Nov. 2006.

[7] 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 Video Retrieval Workshop*, Gaithersburg, MD, Nov. 2005.

[8] T.-S. Chua, S.-Y. Neo, Y. Zheng, H.-K. Goh, Y. Xiao, M. Zhao, S. Tang, S. Gao, X. Zhu, L. Chaisorn, and Q. Sun. TRECVID-2006 by NUS-I^{2}R. In *NIST TRECVID Video Retrieval Workshop*, Gaithersburg, MD, Nov. 2006.

[9] W. B. Croft. Combining approaches to information retrieval. In *Advances in Information Retrieval*. Kluwer Academic Publishers, 2000.

[10] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407, 1990.

[11] T. Dunning. Accurate methods for the statistics of surprise and coincidence. *Comp. Linguistics*, 19(1):61–74, Mar. 1993.

[12] S. Gauch and J. Wang. A corpus analysis approach for automatic query expansion. In *CIKM '97: Proc. 6th Intl. Conf. on Information and Knowledge Management*, pages 278–284, Las Vegas, NV, 1997.

[13] A. Haubold, A. Natsev, and M. Naphade. Semantic multimedia retrieval using lexical query expansion and model-based reranking. In *Intl. Conf. on Multimedia and Expo (ICME)*, Toronto, Canada, July 2006.

[14] A. G. Hauptmann, M. Christel, R. Concescu, J. Gao, Q. Jin, W.-H. Lin, J.-Y. Pan, S. M. Stevens, R. Yan, J. Yang, and Y. Zhang. CMU Informedia's TRECVID 2005 Skirmishes. In *NIST TRECVID Video Retrieval Workshop*, Gaithersburg, MD, Nov. 2005.

[15] W. H. Hsu, L. S. Kennedy, and S. F. Chang. Video search reranking via information bottleneck principle. In *Proc. 14th Annual ACM Intl. Conference on Multimedia*, pages 35–44, Santa Barbara, CA, 2006.

[16] Y. Jing and W. B. Croft. An association thesaurus for information retrieval. In *Proc. RIAO'94*, pages 146–160, 1994.

[17] K. S. Jones. *Automatic Keyword Classification for Information Retrieval*. Butterworths, 1971.

[18] L. Kennedy, A. Natsev, and S.-F. Chang. Automatic discovery of query-class-dependent models for multimodal search. In *ACM Multimedia 2005*, pages 882–891, Singapore, Nov. 2005.

[19] R. Manmatha and H. Sever. A formal approach to score normalization for meta-search. *Proceedings of HLT*, pages 98–103, 2002.

[20] M. Naphade, J. Smith, and F. Souvannavong. On the detection of semantic concepts at TRECVID. In *ACM Multimedia*, New York, NY, Nov 2004.

[21] M. Naphade, J. R. Smith, J. Tesic, S. F. Chang, W. Hsu, L. Kennedy, A. Hauptmann, and J. Curtis. Large-Scale Concept Ontology for Multimedia. *IEEE MultiMedia Magazine*, 13(3):86–91, 2006.

[22] A. Natsev, M. R. Naphade, and J. R. Smith. Semantic representation, search and mining of multimedia content. In *10th ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining (KDD '04)*, Seattle, WA, Aug. 2004.

[23] A. Natsev, M. R. Naphade, and J. Tevsić. Learning the semantics of multimedia queries and concepts from a small number of examples. In *ACM Multimedia 2005*, Singapore, Nov. 6–11 2005.

[24] S.-Y. Neo, J. Zhao, M.-Y. Kan, and T.-S. Chua. Video retrieval using high level features: Exploiting query matching and confidence-based weighting. In *CIVR 2006*, pages 143–152, Tempe, AZ, July 2006.

[25] S. Patwardhan, S. Banerjee, and T. Pedersen. Using measures of semantic relatedness for word sense disambiguation. In *Conference on Intelligent Text Processing and Computational Linguistics*, pages 241–257, Mexico, Feb. 16–22 2003. Springer.

[26] C. Peterson and J. Anderson. A mean field theory learning algorithm for neural networks. *Complex Systems*, 1:995–1019, 1987.

[27] J. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in Large Margin Classifiers*, 10(3), 1999.

[28] Y. Qiu and H. Frei. Concept based query expansion. In *Proc. 16th Annual Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval*, pages 160–169, 1993.

[29] N. Rasiwasia, N. Vasconcelos, and P. J. Moreno. Query by semantic example. In *CIVR 2006*, pages 51–60, Tempe, AZ, July 2006.

[30] P. Resnik. Using Information Content to Evaluate Semantic Similarity in a Taxonomy. In *Intl. Joint Conf. Artificial Intelligence*, pages 448–453, Montreal, Canada, 1995.

[31] J. J. Rocchio. *Relevance feedback in information retrieval*, chapter 14. Relevance Feedback in Information Retrieval, pages 313–323. Prentice-Hall Inc., Englewood Cliffs, NJ, 1971.

[32] H. Schütze and J. O. Pedersen. A cooccurrence-based thesaurus and two applications to information retrieval. In *Intelligent Information Retrieval Systems RIAO'94*, pages 266–274, New York, NY, 1994.

[33] J. R. Smith, M. R. Naphade, and A. Natsev. Multimedia semantic indexing using model vectors. In *IEEE Intl. Conference on Multimedia and Expo (ICME '03)*, Baltimore, MD, July 2003.

[34] C. Snoek, M. Worring, J. van Gemert, J. Geusebroek, and A. Smeulders. The challenge problem for automated detection of 101 semantic concepts in multimedia. In *Proc. 14th Annual ACM Intl. Conference on Multimedia*, pages 421–430, Santa Barbara, CA, 2006.

[35] C. G. M. Snoek, J. C. van Gemert, J. M. Geusebroek, B. Huurnink, D. C. Koelma, G. P. Nguyen, O. D. Rooij, F. J. Seinstra, A. W. M. Smeulders, C. J. Veenman, and M. Worring. The MediaMill TRECVID 2005 Semantic Video Search Engine. In *NIST TRECVID Video Retrieval Workshop*, Gaithersburg, MD, Nov. 2006.

[36] J. Tevsić, A. Natsev, and J. R. Smith. Cluster-based data modeling for semantic video search. In *ACM CIVR 2007*, Amsterdam, The Netherlands, July 2007.

[37] TREC Video Retrieval Evaluation. National Institute of Standards and Technology. http://www-nlpir.nist.gov/projects/trecvid/.

[38] E. M. Vorhees. Query expansion using lexical-semantic relations. In *17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 61–69, Dublin, Ireland, August 1994.

[39] L. Xie, A. Natsev, and J. Tevsić. Dynamic multimodal fusion in video search. In *ICME 2007*, Beijing, China, 2007.

[40] J. Xu and W. B. Croft. Query expansion using local and global document analysis. In *19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 4–11, New York, NY, 18–22 August 1996.

[41] R. Yan, J. Yang, and A. G. Hauptmann. Learning query-class dependent weights in automatic video retrieval. *Proc. 12th Annual ACM Intl. Conf. on Multimedia*, pages 548–555, 2004.

[42] R. Yang. *Probabilistic Models for Combining Multiple Knowledge Sources in Multimedia Retrieval*. PhD thesis, Carnegie Mellon University, 2006.

[43] Y. Yang and J. O. Pedersen. A comparative study on feature selection in text categorization. In *Proc. 14th Intl. Conf. on Machine Learning (ICML'97)*, 1997.