Inf Retrieval (2007) 10:445–484
DOI 10.1007/s10791-007-9031-y

A review of text and image retrieval approaches for broadcast news video

Rong Yan · Alexander G. Hauptmann

Received: 2 February 2007 / Accepted: 23 July 2007 / Published online: 17 August 2007

Abstract The effectiveness of a video retrieval system largely depends on the choice of underlying text and image retrieval components. The unique properties of video collections (e.g., multiple sources, noisy features and temporal relations) suggest we examine the performance of these retrieval methods in such a multimodal environment, and identify the relative importance of the underlying retrieval components. In this paper, we review a variety of text/image retrieval approaches as well as their individual components in the context of broadcast news video. Numerous components of text/image retrieval have been discussed in detail, including retrieval models, text sources, temporal expansion methods, query expansion methods, image features, and similarity measures. For each component, we conduct a series of retrieval experiments on TRECVID video collections to identify their advantages and disadvantages. To provide a more complete coverage of video retrieval, we briefly discuss an emerging approach called concept-based video retrieval, and review strategies for combining multiple retrieval outputs.

Keywords Video retrieval · Text retrieval · Image retrieval · Concept-based retrieval · Fusion · Review

1 Introduction

Recent improvement in processor speeds, network systems and digital storage has led to an explosion in the amount of online video data. It has been estimated that more than 40% of

R. Yan (✉)
Intelligent Information Management Department, IBM TJ Watson Research, 19 Skyline Drive,
Hawthorne, NY 10532, USA
e-mail: yanr@us.ibm.com

A. G. Hauptmann
Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA 15213, USA
e-mail: alex@cs.cmu.edu

Internet traffic involves peer-to-peer swaps of video content (MacWorld 2006). Add to that the growing amount of legitimate content available through companies such as Apple or Google, and the scale of consumer demand for video begins to emerge. With the availability of such vast amount of digital video content, here comes the need for effective management and search. This is reflected in one of the *SIGMM* grand challenges (Rowe and Jain 2004):

> "Making capturing, storing, finding and using digital media an everyday occurrence in our computing environment".

To achieve this, *video retrieval* systems, which aim to search video collections for documents relevant to an information need, offer an important platform to access and manage the vast amount of online video content. These systems usually start by asking users to provide a multimodal query, and then match it with a collection of indexed video documents in a database. The query may consist of only text, only images, or contain multimodal information such as text, image, audio or video examples. The document relevance w.r.t. the query is determined by a set of retrieval status values (RSVs) for the documents. Typical video retrieval systems (Lew et al.2002; Smith et al.2002; Wactlar et al.1999; Zhang et al.1994) represent video documents as a set of low-level detectable features and high-level semantic concepts (Snoek and Worring 2005; Christel and Hauptmann 2005). The retrieval task can then be formulated as a fusion problem that aims to combine a bag of retrieval outputs generated from multiple retrieval sources, such as text retrieval on speech transcripts and image retrieval on color histograms. To illustrate, Fig. 1 shows the design of a typical video retrieval system (Hauptmann et al.2003a) for broadcast news video. First, video footage is segmented into a number of smaller clips as "documents". Various sets of low-level features are extracted from the video clips through analyzing multimedia sources. Each video clip is then associated with a vector of ranking features, which include both individual outputs from different retrieval experts indicating the query-document similarity from a specific modality, as well as the detection outputs of semantic concepts. Finally, the system combines these ranking features based on the query description to produce a final ranked list of video documents.

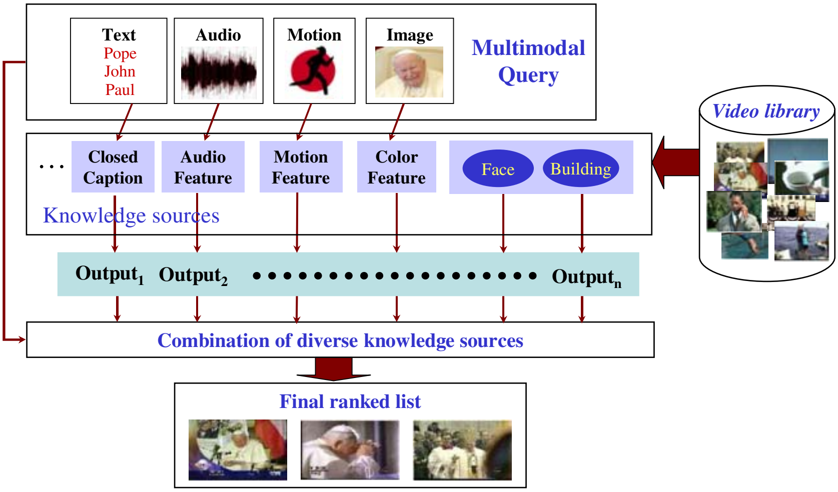


Fig. 1 Design of typical video retrieval systems for broadcast news video

In practice, the choice of text and image retrieval algorithms often has a considerable impact on the final performance of video retrieval. Although both text retrieval and (content-based) image retrieval have been extensively studied on uni-modal collections before (Kraaij 2004; Smeulders et al.2000; Antani et al.2002; Lew et al. 2006), it is still worthwhile to reexamine their effectiveness in a multimodal environment such as broadcast news video. This is because video has unique properties that distinguish it from traditional unimodal data. For instance, in video collections, (1) text features are generated from heterogenous and noisy information sources; (2) a temporal proximity between two video documents might indicate their content similarity; (3) the number of images extracted from video collections can be much larger than that from image collections, but the quality is usually worse; (4) the combination of text and image retrieval can compensate for the weaknesses of each modality; (5) each video document can be associated with multiple levels of representations, including low-level text/visual features and high-level semantic concepts.

In this paper, we present a review of typical text and image retrieval approaches in state-of-the-art video retrieval systems. Several video-specific issues have been extensively investigated, such as the choice of text sources and the temporal expansion methods. In order to provide a fair comparison, we conduct a number of retrieval experiments on TRECVID video data sets to corroborate our conclusions. We also briefly discuss an emerging retrieval approach based on high-level semantic concepts, called concept-based video retrieval, as well as the strategies to combine multiple retrieval outputs from different modalities. Note that we do not aim to completely cover the entire area of video retrieval (only broadcast news video is discussed rather than other genres), neither do we intend to present an exhaustive survey on all the individual components of video retrieval systems (the discussions on several related aspects are left out, such as human-computer interface, data indexing, storage and so forth). Instead, our goal is to present a broad overview of the standard text and image retrieval approaches on broadcast news video, and provide a series of experimental results to justify their advantages and disadvantages.

1. Find shots of Yasser Arafat.
2. Find shots of a rocket or missile taking off.
3. Find shots of the Tomb of the Unknown Soldier at Arlington National Cemetery.
4. Find shots of the front of the White House in the daytime with the fountain running.

Fig. 2 Text query examples from TRECVID 2003

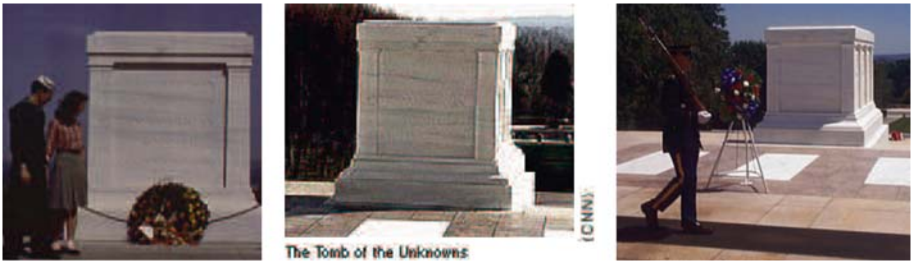


Fig. 3 Image query examples of "Find the Tomb of Unknown Soldiers at Arlington National Cemetery"

The rest of this paper is organized as follows. Section 2 provides a brief review of the TRECVID evaluation and its video collections, which serve as the testbeds for us to design retrieval experiments. Then, the content description is decomposed into three parts. In Sect. 3, we discuss four aspects of text retrieval approaches, i.e., retrieval models, text sources, temporal expansion window and query expansion strategies. In Sect. 4, we discuss image retrieval approaches by image features, distance metrics and query fusion methods. In Sect. 5, we discuss the concept-based video retrieval and multimodal combination strategies. Finally, we conclude the paper in Sect. 6.

2 Brief review on TREC video retrieval evaluation

The National Institute of Standards and Technology (NIST) has sponsored the annual *Text REtrieval Conference* (TREC) to encourage research within the information retrieval community by providing the infrastructure and benchmark necessary for large-scale evaluation of retrieval methodologies. In 2001, NIST started the TREC Video Track (now referred to as TREC Video Retrieval Evaluation, a.k.a. (TRECVID); Smeaton and Over 2003) to promote progress in content-based video retrieval via an open, metrics-based evaluation, where the video corpora have ranged from documentaries, advertising films, technical/educational material to multi-lingual broadcast news. Since 2003 TRECVID spun off as a separate and independent entity, which is co-located with TREC every year. The international participation of TRECVID has rapidly grown from 12 organizations and academic institutions in 2001 to 62 in 2005 (Smeaton et al.2006).

The TRECVID forum defined a number of tasks, including shot detection, story segmentation, semantic feature extraction and automatic/manual/interactive search. Among them, the search tasks in TRECVID are extensions of their text analogues from previous TREC evaluations. Participating groups are required to index a given test collection of video data and return lists of relevant clips. The search topics are designed as multimedia descriptions of an information need, which might contain not only text keywords but also possibly video, audio and image examples. Typically, the topics include requests for some specific items, specific people, specific facts, instances of categories and instances of activities. By analogy to “document“ in text retrieval, TRECVID adopts the basic video units to be retrieved as video shots, which is defined as a single continuous camera operation without cut, fade or dissolve. To simplify ranking, the rank lists contain only up to N shots relevant to the query where N = 100 for 2002, N = 1,000 for 2003 and 2004. The shot boundaries and the ground truth of search results are officially provided. The ground truth is pooled from all the submissions.

The search task distinguishes between (1) interactive approaches, where a user can interact with the system repeatedly to locate the relevant shots, (2) manual approaches, where a user is only allowed to modify the query before submitting it to the retrieval system, and (3) automatic approaches, where systems must directly parse the queries and provide the relevant results. *Precision* and *recall* are two central criteria to evaluate the performance of retrieval algorithms. Precision is the fraction of retrieved documents that are relevant. Recall is the fraction of relevant documents that are retrieved. NIST also defines another measure of retrieval effectiveness called non-interpolated average precision over a set of retrieved documents (shots). Let R be the number of true relevant documents in a set of size S, L be the ranked list of documents, and R_j be the number of relevant documents in the top j documents. Let $I_j =$ 1 if the jth document is relevant and 0 otherwise. The non-interpolated average precision (*AP*) is then defined as $\frac{1}{R}\sum_{j=1}^{S}\frac{R_j}{j}*I_j$. Mean average precision (*MAP*) is the mean of average precisions over all queries.

Providing one of the largest publicly available video collections with manual annotations, TRECVID has become a standard large-scale testbed for the task of video retrieval. Each video collection of TRECVID'03–'05 is split into a development set and a search set chronologically by source. The development sets are used as the training pool to develop automatic indexing/retrieval algorithms in low-level video feature extraction, high-level semantic feature extraction and search tasks. The search sets mainly serve as testbeds for evaluating the performance of retrieval systems. In Table 1, we list the statistics for all the data collections and their abbreviations used in the following experiments. We also provide additional information of these TRECVID video collections as follows

- *TRECVID2002*: In this collection, NIST defined 25 search topics to find within a search test collection of 40.12 h of video from the Prelinger Archives and the Open Video archives. The material consists of advertising, educational, industrial, and amateur films produced between the 1910s and 1970s. The search test collection was delineated into 14,524 video shots as the common shot reference.
- *TRECVID2003*: In this collection, NIST defined another 25 search topics and provided 120 h (240 30 min programs) of ABC World News Tonight and CNN Headline News recorded by the Linguistic Data Consortium from late January through June 1998, along with 13 h of C-SPAN programming mostly from 2001. Among them, 6 h were used for shot boundary detection, and the reminder was split into a development set including 47,531 shots representing 62.2 h of video and a search set including 75,850 shots representing 64.3 h of video. The TRECVID organizers also provided an official keyframe for each shot from this year.
- *TRECVID2004*: In this collection, 25 search topics were defined. NIST provided a new set of approximately 70 h (48,818 shots) of video as the search set, captured by the Linguistic Data Consortium during the last half of 1998 from both CNN Headline News and ABC World News Tonight. The data was used with the TRECVID2003 data as the development set, so that investments in annotating and truthing the news genre can be reused and iteratively improved. The Informedia project (Wactlar et al.1999) provided a large set of low-level features for the 2004 development data as a common reference.
- *TRECVID2005*: In this collection, 24 search topics were defined. The video collection included a 170 h (nearly 150,000 shots) multilingual news video captured from MSNBC (English), NBC Nightly News (English), CNN (English), LBC(Arabic), CCTV(Chinese) and NTDTV (Chinese). Among them, 6 h of video were used for shot

Table 1 Labels of TRECVID video collections and their statistics

Collection	No. of queries	No. of shots	Duration (h)	Explanation
t02s	25	24,263	40.12	TREC'02 search set
t03d	25	47,531	62.20	TREC'03 development set
t03s	25	75,850	64.30	TREC'03 search set
t04d	24	124,097	127.00	TREC'04 development set
t04s	24	48,818	70.00	TREC'04 search set
t05d	24	74,532	80.00	TREC'05 development set
t05s	24	77,979	80.00	TREC'05 search set

boundary detection, and the reminder was split into a development set including 74,532 shots representing 80 h of video and a search set including 77,979 shots representing another 80 h of video.

3 Text retrieval

As one of the most important retrieval components for video retrieval, text retrieval aims to retrieve a number of top-ranked video documents based on the similarity between query keywords and text features of documents. Textual features can be extracted from a number of information sources such as speech transcripts, closed captions and video optical character recognition. They are usually the least complex features to process and analyze. They are also reliable in handling semantic queries for broadcast news video (Hauptmann and Christel 2004). However, textual features also suffer from several drawbacks. For instance, they are noise-prone due to feature generation errors, and they do not always convey visual content on the screen, especially for dialogue-based genres such as movies, sitcoms and documentaries. In this section, we describe several key components of text retrieval, including retrieval models, text sources, temporal expansion strategies and query expansion approaches. For each component, we discuss the strengthes/weaknesses of its possible configurations and evaluate its retrieval performance using the TRECVID collections.

3.1 Retrieval Models

The state-of-the-art text retrieval algorithms typically fall into one of the following two categories, i.e., vector space models and probabilistic models. In vector space models, relevance between a document d and a query q is defined based on a distance measure on a high-dimensional term vector space. In contrast, classical relevance-based probabilistic models consider the difference of term distribution between relevant documents and non-relevant documents. Beyond relevance-based probabilistic models, language-model based probabilistic retrieval models assume that a query is generated by combining a language model estimated on the current document and a smoothed model estimated on the entire collection. Previous experiments (Ponte and Croft 1998) have shown that vector space models and probabilistic retrieval models perform roughly on par with each other. In this section, instead of investigating all possible retrieval models, we only discuss the most representative models and use the standard parameters to evaluate them in video collections. A more complete survey of text retrieval models can be found in (Kraaij 2004).

3.1.1 Vector space model

Formally, vector space models represent each document D_k (the kth document) and query Q as a vector w.r.t. indexed terms, $D_k = [d_{k0}, d_{k1}, .., d_{kW}]$, $Q = [q_0, q_1, .., q_W]$, where d_{ki} is the weight associated with the ith indexed term for D_k and q_i is the weight associated with the ith term for Q. Typically the term weights are very sparse, and thus most weights are equal to zero. The query-document similarity can be computed using the inner product between their term weights, $\mathrm{sim}(Q, D_k) = \sum_{i=1}^{W} q_i d_{ki}$. The simplest approach to design term weights is to associate each term with a binary value {0, 1} according to the term presence. However, term weights do not need to be binary. Instead, term weights can be any positive real values that encode distribution information of the indexed words. For example, one of the most popular term weighting schemes is called *tf.idf* (Salton 1989), which suggests term weights are proportional to the frequency of the term occurrence within a document, and inversely proportional to the number of documents where the terms appear. Usually, the *idf* term is converted to its logarithmic value to make it comparable to the *tf* term. By multiplying these two factors together, we can derive the following similarity measure,

$$\mathrm{sim}(Q, D_k) = \sum_{i=1}^{W} q_i^{tf} \cdot q_i^{idf} \cdot d_{ki}^{tf} \cdot d_{ki}^{idf}.$$

Another dimension of designing term weights is the document length normalization scheme that attempts to eliminate the effects of heterogeneous length distribution of text documents. They are useful because if no document length normalization is applied, the retrieval results tend to be biased to long documents that contain more content words to match. The most well-known document length normalization schemes are the Cosine similarity normalization (Buckley and Walz 1999) and the pivoted document length normalization (Singhal et al.1996). The Cosine similarity normalization divides the term weights of kth document by a factor of $\sqrt{\sum_i d_{ki}^2}$ so that the sum of squared term weights are normalized to 1. Despite the intuitive explanation, it was found that the Cosine normalization scheme was not always optimal in practical datasets like TREC collections (Buckley and Walz 1999). One problem is that long documents in these collections contain many unique terms with misspellings and thus they become unretrievable due to the low term weights. As a better alternative, pivoted normalization proposed by Singhal et al. (1996) intends to normalize the document vectors by a factor of $(1-s)p + sV_d$, where V_d is the document length and two other parameters are slope s and pivot p. It is designed in a way to boost the retrieval scores of short documents and decrease the scores of long documents. The parameters of pivoted normalization can be pre-defined or learned on a previous collection. This type of normalization has been proved to be successful in practice, although the motivation is not as elegant as Cosine normalization.

3.1.2 Probabilistic models and okapi models

In contrast to vector space models which have a less elegant theoretical basis, probabilistic models provide a more principled framework by translating information retrieval into an uncertainty inference problem. The underlying principle for probabilistic models is called the *Principal ranking principle* (Robertson 1977), which suggests sorting documents by the log-odds of relevance. If we define R as the binary relevance variable where $R = 1$ means the document D is relevant to Q and vice versa, the log-odds ratio can be defined as $\log \frac{P(R=1|D,Q)}{P(R=-1|D,Q)}$. Given this general principle, we can categorize probabilistic models into three classes,

- *Probabilistic relevance model* (Robertson and Sparck Jones 1977; Robertson et al.1992): document relevance is estimated given the distributions of indexed terms in relevant documents and irrelevant documents;

- *Inference based model* (Turtle 1991): the retrieval problem is formulated as a Bayesian inference network;
- *Language-model based model* (Ponte and Croft 1998; Zhai and Lafferty 2001): the query is generated by a statistical language model on the given document.

The simplest probabilistic relevance model is the binary independence retrieval (BIR) model (Robertson and Sparck Jones 1977). In this model, each document is represented as a binary vector of the term presence/absence where all the information of term frequencies is discarded. To estimate the term weights, BIR proceeds by inverting the position of R and D based on Bayes rule and estimating the generative probabilities of document D in relevant and irrelevant documents. But the BIR model has its own limitations, e.g., it only considers the binary presence of indexed terms with frequency information discarded. To address this, Robertson and Walker (1994) considered a series of approximations to a new probabilistic model based on two Poisson distributions and finally proposed a series of retrieval models named the Okapi models. Among them, the best known Okapi model is the BM25 version (Robertson et al.1992),

$$\sum_{i=1}^{N_q}\left(s_1 s_3 \cdot \frac{d_{ki}^{tf}}{d_{ki}^{tf}+K^c} \cdot \log\frac{N-n+0.5}{n+0.5} \cdot \frac{q_i^{tf}}{q_i^{tf}+k_3}\right) + k_2 \cdot |Q| \frac{\Delta - d}{\Delta + d}, \tag{1}$$

where $K^c = k_1((1-b) + bd/\Delta)$, k_1, k_2, k_3, s_1, s_3, b are the predefined constants that have to be decided empirically, d is the document length, Δ is the average document length, $|Q|$ is the number of query terms, n is the number of documents containing the query term q_i and N is the size of document collections. It can be found that several tuning parameters in Eq. 1 are needed to be determined empirically. Researchers have come up with a simplified retrieval model as follows by setting $k_2 = 0$, $b = 0.75$, $k_1 = 2$, $s_1 = 1$, $s_3 = k_3 + 1$, $k_3 = \infty$ and leaving out the document length correction component,

$$\sum_{i=1}^{N_q}\left(q_i^{tf} \cdot \frac{d_{ki}^{tf}}{d_{ki}^{tf} + 2 \times (0.25 + 0.75 \times \frac{d}{\Delta})} \cdot \log\frac{N-n+0.5}{n+0.5}\right), \tag{2}$$

this variant, also known as the SMART version of BM25 model (Kraaij 2004), has been widely applied in previous work. The document length normalization in this variant reflect the intuition that the longer the text document, the greater the likelihood that a particular query term will occur by chance. More details about the Okapi retrieval models can be found at (Robertson et al.1992).

Recently, statistical language modeling approaches (Ponte and Croft 1998; Zhai and Lafferty 2001) have emerged as a new probabilistic model for information retrieval, which stems from its earlier counterparts in the field of speech recognition. In this approach, each document is associated with a language model which is a probability distribution over terms. The document ranking is determined by the conditional probability of the query given the language model of documents. Formally, the conditional probability of query Q given a document D is defined as, $P(Q|D) = P(q_1, q_2, \ldots, q_n|D) = \prod_{t=1}^{N} P(q_j|D)$, where the last step is derived based on the independence assumption of query terms given the document language model. Typically, language models from documents could be built efficiently and its performance is demonstrated to be on par with the vector space model. Although lack of the notion of relevance in the language modeling approaches is a setback for some applications such as relevance feedback, Lafferty and Zhai (2003) showed that relevance can be represented as an implicit variable in modeling and thus what we are actually estimating is $P(Q|D,R)$. To put another way, language modeling and relevance-based probabilistic models are actually two sides of the same coin. Since language modeling retrieval approaches provide such a formal framework for information retrieval, they have been successfully applied in several other retrieval tasks. For example, these approaches have gained its success in multimedia retrieval (Westerveld 2004; Iyengar et al.2005) by jointly modeling text features with multinomial distributions and image features with mixture of Gaussian distributions.

In the following experiments, we choose retrieval models from the Okapi family (Robertson et al.1992) to represent the entire set of probabilistic retrieval models, because they have been proved to effective in text retrieval on a variety of data collections. Besides, Okapi models have been adopted in most of the state-of-the-art video retrieval systems (Hauptmann et al.2004; Amir et al.2003; Chua et al.2004; Adcock et al.2004; Snoek et al.2004; Rautiainen et al.2004a; Kennedy et al.2005; Gaughan et al.2003) However, we also realized that statistical language modeling approaches have recently become more popular due to its elegant statistical foundation and comparable performance with the Okapi function (Lafferty and Zhai 2003; Westerveld 2004; Cooke et al.2004; Srikanth et al.2005). But due to space limit, we leave the evaluation of language-model based approaches to future work.

3.1.3 Experiments

To evaluate the performance of various retrieval models, we designed a series of experiments on the TRECVID ’02–’05 collections. For each query topic, the relevance judgment on search sets was officially provided by NIST and the judgment on development sets was collaboratively collected by several human annotators using the Informedia client (Hauptmann et al.2004). All available text sources in the video collections were indexed to be searchable by the retrieval algorithms. More details of these text sources can be found in the next section.

In each collection, text query keywords are automatically extracted by extracting the noun phrases from the original TRECVID query description. In order to compare the retrieval performance in a finer granularity, we manually assigned each query into one of the five query classes, that is, named person, named objects, general objects, sports and others. As a preprocessing step, frequent words from a stopword list were removed from both documents and queries. The Porter stemming algorithm was applied to remove morphological variants. Moreover, because the temporal proximity in the video collection is a strong hint to indicate semantic content closeness (Hauptmann and Christel 2004), the retrieval scores were also propagated to a number of nearby documents so as to capture temporal relations within neighbor shots.

As shown in Table 2, we implemented five retrieval methods based on both the vector space models and the probabilistic models. Two of them belong to the variants of BM-25 Okapi models which includes the aforementioned SMART version BM-25 model (**Okapi-SM**) and a BM-25 model with different parameters setting $k_1 = 1$, $b = 0.5$ in order to reduce the effect of document length normalization (**Okapi-LM**). The other three models are based on the vector space models that can be represented using the six-digit SMART codes (Buckley and Walz 1999), i.e., **lnn.ntn** with idf and log-tf weights, **nnp.ntp** with idf weights and pivot length normalization where $s = 0.2$ and p is average document length and **nnn.ntn** with idf weights. For each retrieval method, we reported their mean average

Table 2 Comparison of text retrieval models

Data	Method	MAP	P30	P100	Person	S-Obj	G-Obj	Sports	Others
t05s	Okapi-SM	0.069	0.174	0.162	0.144	0.058	0.016	0.069	0.017
	Okapi-LM	0.067	0.171	0.160	0.141	0.056	0.016	0.064	0.016
	nnn.ntn	**0.074**	0.172	0.162	0.159	0.057	0.016	0.081	0.016
	lnn.ntn	0.070	0.174	0.161	0.149	0.057	0.016	0.072	0.016
	nnp.ntp	**0.074**	0.174	0.164	0.158	0.057	0.016	0.083	0.016
t04s	Okapi-SM	0.078	0.178	0.105	0.189	0.000	0.039	0.047	0.041
	Okapi-LM	**0.079**	0.180	0.105	0.192	0.000	0.039	0.047	0.042
	nnn.ntn	0.072	0.154	0.100	0.169	0.000	0.039	0.043	0.040
	lnn.ntn	0.078	0.171	0.104	0.186	0.000	0.040	0.046	0.041
	nnp.ntp	0.071	0.152	0.097	0.170	0.000	0.038	0.042	0.035
t03s	Okapi-SM	0.150	0.184	0.120	0.372	0.237	0.067	0.033	0.007
	Okapi-LM	**0.151**	0.187	0.120	0.375	0.245	0.066	0.032	0.008
	nnn.ntn	0.123	0.168	0.111	0.262	0.217	0.065	0.035	0.006
	lnn.ntn	0.142	0.183	0.119	0.339	0.235	0.064	0.037	0.007
	nnp.ntp	0.122	0.168	0.111	0.259	0.219	0.063	0.033	0.006
t02s	Okapi-SM	**0.108**	0.109	0.075	0.175	0.184	0.082	0.000	0.011
	Okapi-LM	0.107	0.113	0.074	0.174	0.184	0.081	0.000	0.011
	nnn.ntn	0.101	0.109	0.073	0.135	0.180	0.085	0.000	0.008
	lnn.ntn	0.101	0.109	0.074	0.134	0.182	0.082	0.000	0.010
	nnp.ntp	0.097	0.101	0.069	0.120	0.180	0.081	0.000	0.009
t05d	Okapi-SM	0.032	0.065	0.058	0.077	0.015	0.006	0.022	0.009
	Okapi-LM	0.031	0.064	0.057	0.073	0.017	0.010	0.021	0.009
	nnn.ntn	**0.036**	0.082	0.061	0.093	0.013	0.004	0.026	0.009
	lnn.ntn	0.033	0.069	0.058	0.082	0.014	0.004	0.023	0.009
	nnp.ntp	**0.036**	0.082	0.062	0.093	0.013	0.004	0.027	0.009
t04d	Okapi-SM	0.073	0.097	0.075	0.130	0.000	0.061	0.077	0.036
	Okapi-LM	**0.074**	0.101	0.075	0.136	0.000	0.068	0.073	0.031
	nnn.ntn	0.065	0.088	0.075	0.116	0.000	0.050	0.063	0.042
	lnn.ntn	0.072	0.092	0.073	0.121	0.000	0.057	0.076	0.045
	nnp.ntp	0.066	0.076	0.075	0.124	0.000	0.050	0.064	0.037
t03d	Okapi-SM	0.092	0.077	0.051	0.185	0.102	0.080	0.048	0.009
	Okapi-LM	**0.095**	0.077	0.053	0.190	0.112	0.082	0.049	0.009
	nnn.ntn	0.078	0.049	0.045	0.172	0.038	0.085	0.047	0.009
	lnn.ntn	0.083	0.069	0.049	0.178	0.065	0.080	0.049	0.009
	nnp.ntp	0.076	0.051	0.046	0.172	0.033	0.084	0.047	0.009

The model(s) with the highest MAP is bolded for each collection

precision (MAP), mean precision at 30 documents (P30) and mean precision at 100 documents (P100) over all queries. The MAP of each query type are also reported.

By comparing these five retrieval approaches, we can observe that both Okapi models outperform the vector space models in almost all the cases w.r.t. MAP, Prec30 and Prec100, especially on the TRECVID’02-’04 collections. This observation is consistent with previous TREC ad-hoc retrieval results (Robertson et al.1992) which demonstrated the effectiveness of Okapi models in large scale text collections. Moreover, out of the three vector space models, the one using logarithm term frequency and *idf* weights produces a comparable performance with the Okapi retrieval models. In fact, the Okapi models and the *log-tf* model share a common property that term frequency has relatively lower influences in the retrieval function as compared to "raw" term frequency in other models. This shows the usefulness of normalizing term frequency to a certain range. It also partially explains the popularity of selecting a logarithm *tf* weighting scheme in the vector space models. However, there are some exceptions where the Okapi and *log-tf* models do not work well, i.e., the t05d/t05s collections, where several closely related text sources are generated using multiple versions of speech recognizers and machine translators provided by NIST, Microsoft, Sphinx and CMU ISL (Hauptmann et al.2005). In this case, the raw term frequency *tf* turns out to be a more useful factor with high recognition/translation errors but multiple complementary sources. But it is worth pointing out that the differences between these retrieval methods are not statistically significant[1]. More experiments are needed to further clarify the advantages and disadvantages between retrieval models.

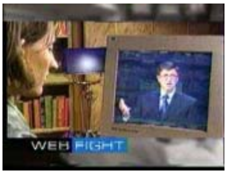
Along another line, we notice that the retrieval performance is relatively insensitive to the choice of document length normalization schemes. This is due to the fact that text length in video documents is relatively stable in contrast to regular text documents, which can have very skew length distribution. Finally, the last couple of columns allows us to further analyze the retrieval models across query types. Roughly speaking, text retrieval methods are most effective for the queries of finding persons and sometimes finding specific objects, because in these cases the information needs are clearly defined as terms to be retrieved, and thus the relevant shots will be around the associated texts, as long as the words are mentioned in the video. However, the other three types of queries gain less benefit from text retrieval, because their information needs are usually less related to text keywords. For example, the information need of "finding a roundtable meeting with a large table" would be much more difficult to be captured by text than that of "finding Hu Jintao, President of China". For these query types, we should consider incorporating other modalities to improve retrieval performance.

3.2 Text sources

Text data in video collection are not always generated from a single source. Instead, a lot of video corpora, such as broadcast news, can be associated with multiple text sources extracted via either manual annotation or some well-established automatic techniques, such as audio signal processing or visual appearance analysis. Given a large number of text sources available, it is interesting to study what are distinctive properties for each type of text information and which text source can contribute most to video retrieval. Generally speaking, the text sources span several dimensions as follows: (Smeaton and Over 2003)

- Closed captions (CC) which contain the accurate spoken text written by a person;
- Automatic speech transcripts (ASR) which are converted from raw audio signals by speech recognizers;
- Video optical character recognition (VOCR) extracted from the text visible in the screen;
- Production metadata such as titles, published descriptions of the video and audio description for movies.

[1] In this paper, we compute the *p*-value using a sign test and set the significant level to be 1%.



VOCR: WeE! Fiighht
ASR: Microsoft the rest of the. Should we be afraid of this computer? is there reason to be great?
CC: Microsoft and the rest of us. Should we be afraid of this computer giant? is there reason to be grateful?

Fig. 4 Examples of three different types of text sources: VOCR, ASR, and CC

Figure 4 shows some examples of the text sources mentioned above. Unlike traditional text collections, most text sources from video are more or less noisy because of either human annotation errors or automatic processing mistakes. Among them, CC is the most complete and accurate source with the lowest word error rate. Unfortunately, CC is not always available for retrieval, unless the video collections have been manually transcribed or captioned with keywords before the retrieval process. Although a considerable fraction of the television broadcasts have manual transcription nowadays, a much larger number of video collections are unfortunately not transcribed because of the high cost of human transcription (Hauptmann 2006).

When CC is not available, ASR is often extracted as an important supplementary text source, which is obtainable through automatic speech recognition (Huang et al.1993; Gauvain et al.2002). Although ASR might have a large number of recognition errors, previous experiments (Hauptmann 2006) showed that as long as speech recognition has a word error rate lower than 35%, the retrieval performance using spoken documents is only 3–10% worse than that using perfect text transcripts. Moreover, at times the speech transcripts can be improved based on evidence from other modalities. For instance, Yang et al. (2003) attempted to correct non-English names by matching them with VOCR text. It is also worth pointing out that even when CC is available, the retrieval system might still need to consult the corresponding time alignment information from ASR to synchronize CC with the spoken words. All of these factors have made speech transcripts one of the most indispensable text sources in video retrieval systems.

Another textual feature can be derived visually by extracting the overlayed text presented in the video images via video optical character recognition (VOCR) techniques (Hua et al.2001; Sato et al. 1998; Chen and Odobez 2005). VOCR is a useful tool to capture people names, event names, location names, as well as product names in commercials that are sometimes not explicitly mentioned in the transcript. A complete survey of VOCR related approaches can be found at (Lienhart 2003). VOCR technologies have been commercially available for a long time, but unfortunately their output often exhibits a high error rate. For example, the word accuracy of VOCR on the TRECVID'01 video collection is estimated to be as low as 27% (Hauptmann et al.2003b). To address this issue, several text correction methods have been proposed to post-process and correct the errors. Among them, dictionary-based correction (Hauptmann et al. 2003b) that expands VOCR words into its other possible spelling based on a dictionary such as MS Word has been demonstrated to be an effective correction approach.

3.2.1 Experiments

In this section, we evaluate five different configurations of the text sources in order to explore the characteristics of each text source for video retrieval. The Okapi-SM retrieval model is used as the baseline retrieval method. Not all the text sources are available in every video collection. In the collection *t*02*s*, the available sources are ASR, VOCR and production metadata. In the other collections, no production metadata is officially provided, but CC are obtainable in the TRECVID'03–'04 collections. In the following experiments, CC and production metadata are officially provided by NIST. The ASR is a mixture of the NIST-provided transcripts and the outputs from a large vocabulary, speaker independent speech recognizer with the word error rate around 30% (Huang et al.1993). The VOCR transcript is generated by commercial OCR software which process the filtered images of alphanumeric symbol into text. The screen text is further improved through the use of dictionaries and thesauri (Hauptmann et al.2003a).

Table 3 compares the text retrieval performance on various sources. As expected, if only one text source is chosen, the CC usually provides the best retrieval results due to its low transcription error rate. However, by comparing mean average precision based on ASR and CC, we can find that their performance are roughly on par with each other (with a difference around 2% w.r.t. MAP), which justify the use of ASR even if they come with a lot of mis-recognition errors. Also, it is not surprising to observe that VOCR produces the worst MAP among all text sources, because it has a high recognition error and the textual information represented in the screen is often too concise for retrieval purpose. VOCR tends to be more useful for the queries of finding persons than other query types due to the frequent appearance of person names on the screen. The production metadata provided in TREC'02 is useful but less effective than ASR, because they contain less useful semantic content in the shot level. Finally, we found that combining two or more different text sources almost always achieves a higher performance than using any of the single sources because they generally contain complementary information. For example, in the collection *t*03*s* the configuration that combines all three sources (CC, ASR, and VOCR) together brings a 3.1% MAP improvement over the configuration using closed captions alone. The only exception is in the collection *t*02*s*, where the performance drops significantly when VOCR adds in. This is because the unexpected low performance of VOCR introduces too much noisy information and thus dilutes the better retrieval outputs offered by other sources.

3.3 Temporal expansion window

In contrast to traditional text collections, video collections have a very distinctive property which might greatly compromise the retrieval performance based on textual features, i.e., the misalignments between relevant video shots and relevant keywords in the corpus. This timing inconsistency between visual appearance and text keywords can partially be explained by the "grammar" in the video production, where the narrative text is designed to introduce or summarize the nearby events shown in a temporal proximity. For example, in TV news video, an anchorperson or a reporter might summarize the news at the beginning followed by the shots of news events and important persons, resulting in a major time offset between the keywords and the relevant video clips. There are also cases where words may be spoken in the transcript with no associated video clips present, e.g., a news anchor might discuss a story for which no video clips are available. This issue will become more serious when we are dealing with ASR and CC rather than VOCR. Based on the statistics provided by Yang et al. (2004), in more than half the cases the relevant shots do not show up in the same shot where the query keywords are mentioned, but before or after the shot.

Table 3 Comparison of text sources

Data	SRC	MAP	P30	P100	Person	S-Obj	G-Obj	Sports	Others
t05s	A	0.064	0.161	0.158	0.137	0.053	0.015	0.055	0.016
	V	0.029	0.103	0.063	0.053	0.036	0.001	0.044	0.003
	A,V	**0.069**	0.174	0.162	0.144	0.058	0.016	0.069	0.017
t04s	C	0.073	0.175	0.105	0.158	0.000	0.046	0.056	0.038
	A	0.050	0.120	0.080	0.121	0.000	0.022	0.035	0.025
	C,A	0.073	0.167	0.101	0.165	0.000	0.039	0.050	0.044
	V	0.019	0.055	0.022	0.065	0.000	0.002	0.000	0.003
	C,A,V	**0.078**	0.178	0.105	0.189	0.000	0.039	0.047	0.041
t03s	C	0.117	0.176	0.112	0.263	0.176	0.069	0.039	0.007
	A	0.103	0.157	0.106	0.157	0.238	0.060	0.022	0.005
	C,A	0.118	0.177	0.113	0.236	0.211	0.067	0.034	0.007
	V	0.051	0.033	0.018	0.164	0.075	0.007	0.000	0.005
	C,A,V	**0.150**	0.184	0.120	0.372	0.237	0.067	0.033	0.007
t02s	A	0.141	0.135	0.072	0.272	0.177	0.126	0.000	0.009
	P	0.105	0.124	0.080	0.132	0.170	0.102	0.000	0.009
	A,P	**0.141**	0.155	0.088	0.265	0.184	0.128	0.000	0.010
	V	0.002	0.012	0.007	0.002	0.001	0.001	0.000	0.003
	A,P,V	0.108	0.109	0.075	0.175	0.184	0.082	0.000	0.011
t05d	A	0.030	0.067	0.058	0.075	0.009	0.007	0.022	0.009
	V	0.011	0.029	0.018	0.012	0.037	0.000	0.002	0.000
	A,V	**0.032**	0.065	0.058	0.077	0.015	0.006	0.022	0.009
t04d	C	0.071	0.086	0.070	0.120	0.000	0.061	0.068	0.045
	A	0.057	0.093	0.066	0.103	0.000	0.062	0.033	0.027
	C,A	0.074	0.092	0.073	0.125	0.000	0.062	0.077	0.043
	V	0.009	0.037	0.015	0.031	0.000	0.005	0.000	0.000
	C,A,V	**0.074**	0.097	0.075	0.130	0.000	0.061	0.077	0.036
t03d	C	0.069	0.067	0.046	0.083	0.021	0.119	0.050	0.009
	A	0.051	0.059	0.040	0.093	0.028	0.063	0.043	0.006
	C,A	0.068	0.071	0.048	0.110	0.031	0.096	0.048	0.009
	V	0.043	0.027	0.015	0.101	0.079	0.016	0.011	0.000
	C,A,V	**0.092**	0.077	0.051	0.185	0.102	0.080	0.048	0.009

The second column indicates the available text sources. A, automatic speech transcript; C, closed caption; P, production metadata; V, video OCR

Generally speaking, there are no simple patterns that can accurately detect such kinds of timing misalignments, but it is arguable that the correct shots are likely to appear in the temporal proximity of the locations where query keywords are mentioned (Hauptmann and Christel 2004). This claim can be confirmed by the successes of extant temporal video browsers that display video shots to users in a temporal order (Christel and Martin 1998; Rautiainen et al.2004b; Snoek et al.2004). Therefore, we can make a legitimate assumption that the closer the shot is to the keyword occurrences, the more possible it has the correct visual appearance. Under this assumption, one common solution to overcome the misalignment problem is to pose an "bell-shape" temporal expansion window on top of the text retrieval results, i.e., add the text retrieval scores of a shot to its nearby shots multiplied by some discount factors α, which is monotonically decreasing with a larger shot-keyword distance. The effectiveness of such a temporal expansion treatment has been demonstrated in many previous studies (Yang et al.2004; Rautiainen et al.2004a; Foley et al.2005). In the literature, there exist multiple options for choosing the shape of discount factors, such as using a manual pre-defined windowing function (Foley et al.2005), a Gaussian distribution function (Yang et al.2004), an exponential distribution function (Rautiainen et al. 2004a) and an absolute discount which fixes the discount factor α to the inverse absolute of the shot distance d_s+1 ($\alpha_s = 1/|d_s + 1|$).

3.3.1 Experiments

Two factors are needed to decide in a temporal expansion process, i.e., the form of discount factors and the temporal expansion window size. But to avoid explosive combinations of experimental configurations, we specifically adopt the absolute discounting schemes in the following discussions and leave the evaluations of other discount factors in the future work. Therefore, we only have to set the expansion window size that controls how many shots before and after the retrieved shots are expanded by adding the additional discounted retrieval scores.

The effect of varying expansion window size is compared in Table 4. The following window sizes have been chosen in our experiments, i.e., 0, 1, 2, 5, 10, and 15. The third column in Table 4 shows that the retrieval performance in terms of mean average precision can always become higher with a larger expansion size and the performance improvement is statistically significant. The improvement is much more considerable when the expansion window size is less than 5, while the performance is gradually saturated to an asymptotic level afterwards. Not surprisingly, the major performance growth factor for a larger window size can be traced to a higher overall recall rate brought by more shots expanded (e.g., the recall at 1,000 shots in t05s grows from 14.5 to 31.2% with an expansion size of 5 shots), while the change in the overall precision is not as obvious as recall in these collections.

Given that most of our collections are news video archives, we also compare the following two settings: allow the shots outsides the same story boundaries to be expanded (**No limit**) and not allow them to be expanded (**Limit to story boundary**)[2]. This comparison could give us some evidence if the misalignments between relevant video clips and relevant keywords would go beyond news story boundaries. As shown in the Table 4, limiting the expansion within story boundaries **"Limit to story boundary"** almost always produces better MAP than the other case **"No limit"**. Even worse, sometimes expanding more video shots in the setting of **"No limit"** might compromise the retrieval results such as in collection *t*03*s*. This series of experiments suggest that we should not expand retrieval results to the video shots outside the story where the relevant keywords are found.

3.4 Query expansion

Most users begin their retrieval process without knowing detailed collection information and retrieval environment. At first, they usually find it difficult to formulate a query that

[2] A news story is defined as a segment of a news broadcast with a coherent news focus which contains at least two independent, declarative clauses (Smeaton and Over 2003)

Table 4 Comparison of expansion window sizes

Data	Size	Limit to story boundary				No limits			
		MAP	P30	P100	R1000	MAP	P30	P100	R1000
t05s	0	0.024(+0%)	0.144	0.103	0.145	0.024(+0%)	0.144	0.103	0.145
	2	0.047(+95%)	0.156	0.143	0.267	0.044(+83%)	0.144	0.133	0.268
	5	0.060(+153%)	0.171	0.154	0.312	0.055(+130%)	0.150	0.142	0.319
	10	0.069(+188%)	0.174	0.162	0.340	0.061(+154%)	0.158	0.149	0.349
	15	**0.072**(+201%)	0.172	0.163	0.353	**0.063**(+161%)	0.158	0.151	0.356
t04s	0	0.055(+0%)	0.141	0.088	0.243	0.055(+0%)	0.141	0.088	0.243
	2	0.071(+29%)	0.152	0.103	0.333	0.071(+29%)	0.154	0.103	0.336
	5	0.076(+39%)	0.170	0.105	0.360	0.077(+41%)	0.172	0.105	0.367
	10	0.078(+43%)	0.178	0.105	0.360	0.079(+45%)	0.180	0.103	0.369
	15	**0.079**(+44%)	0.184	0.107	0.368	**0.080**(+45%)	0.183	0.107	0.362
t03s	0	0.102(+0%)	0.165	0.081	0.269	**0.102**(+0%)	0.165	0.081	0.269
	2	0.135(+32%)	0.181	0.108	0.389	0.091(−10%)	0.148	0.080	0.295
	5	0.144(+41%)	0.185	0.119	0.445	0.096(−5%)	0.137	0.081	0.343
	10	0.150(+47%)	0.184	0.120	0.473	0.099(−2%)	0.141	0.080	0.385
	15	**0.150**(+47%)	0.185	0.120	0.474	0.101(−0%)	0.139	0.083	0.409
t02s	0	0.061(+0%)	0.100	0.055	0.303	0.061(+0%)	0.100	0.055	0.303
	2	0.093(+51%)	0.117	0.076	0.418	0.094(+53%)	0.112	0.072	0.436
	5	0.104(+68%)	0.107	0.076	0.454	0.098(+60%)	0.097	0.075	0.473
	10	**0.108**(+75%)	0.109	0.075	0.475	**0.101**(+63%)	0.087	0.072	0.486
t05d	0	0.012(+0%)	0.051	0.037	0.124	0.012(+0%)	0.051	0.037	0.124
	2	0.021(+82%)	0.063	0.047	0.242	0.020(+67%)	0.060	0.046	0.245
	5	0.028(+137%)	0.067	0.056	0.307	0.025(+113%)	0.058	0.049	0.314
	10	0.032(+170%)	0.065	0.058	0.344	0.027(+131%)	0.060	0.053	0.346
	15	**0.033**(+177%)	0.065	0.058	0.346	**0.027**(+132%)	0.060	0.053	0.329
t04d	0	0.040(+0%)	0.086	0.063	0.311	0.040(+0%)	0.086	0.063	0.311
	2	0.059(+49%)	0.085	0.072	0.429	0.041(+2%)	0.075	0.057	0.366
	5	0.068(+70%)	0.092	0.071	0.479	0.045(+12%)	0.076	0.056	0.413
	10	**0.073**(+83%)	0.097	0.075	0.487	0.048(+21%)	0.079	0.058	0.434
	15	**0.073**(+82%)	0.100	0.075	0.521	**0.049**(+23%)	0.083	0.058	0.459
t03d	0	0.070(+0%)	0.077	0.040	0.258	0.070(+0%)	0.077	0.040	0.258
	2	0.085(+21%)	0.080	0.047	0.377	0.084(+19%)	0.079	0.047	0.373
	5	0.087(+23%)	0.080	0.050	0.458	0.084(+19%)	0.077	0.049	0.425
	10	**0.092**(+30%)	0.077	0.051	0.480	**0.089**(+26%)	0.079	0.052	0.463
	15	**0.092**(+30%)	0.076	0.051	0.481	**0.089**(+26%)	0.076	0.052	0.469

well satisfies their information need, and then they will iteratively refine the queries until the retrieval purpose is achieved. This suggests us to consider an iterative query refinement process that can re-construct the query representation in the hope of retrieving additional useful documents. Due to the problem of short document length and high transcription errors in the text sources of video collections, query reformulation becomes more useful because they can capture the exact information needs and introduce additional retrieval information from users or external sources. Two basic query reformulation approaches are available, i.e., query expansion that expands new terms to the original query and term reweighting that modifies the query term weights (Baeza-Yates and Ribeiro-Neto 1999). In this section, we particularly focus on examining the effects of query expansion methods, which can be grouped into three categories as follows (Baeza-Yates and Ribeiro-Neto 1999):

1. Methods based on manual adjustment or relevance feedback from users,
2. Methods based on the set of documents initially retrieved,
3. Methods based on global information from the entire document collection.

The most straightforward query expansion approach is to directly ask users to modify the queries in each iteration. Users can provide additional keywords or substitute previous keywords after reviewing initial retrieved documents. However, this process usually requires intensive effort from users to come up with appropriate modifications. A more user-friendly approach is to utilize relevance feedback to update query keywords. It begins with asking users to label the relevant documents from an initial list of documents, extract terms from the relevant documents and append additional terms to the query. One of the earliest relevance feedback algorithms was proposed by Rocchio (1971). The feedback iterations modify the query vectors by iteratively increasing the weights of terms contained in positive documents and penalizing the terms in negative documents. Besides the explicit feedback, White et al. (2006) consider the form of implicit feedback which monitors searcher interaction with different representations of top-ranked documents and chooses new retrieval strategies accordingly.

Unfortunately, relevance feedback requires additional manual inputs. Sometimes it is more reasonable to expand queries in an automatic manner. To achieve this, we can adopt (1) either a local strategy that explores information from initially retrieved documents, (2) or a global strategy that analyzes document statistics based on the entire collection. The essence of local strategies, a.k.a. pseudo-relevance feedback (PRF), is to utilize top retrieved documents as positive examples to select discriminative query terms. In practice, the idea of local analysis has been implemented in various forms for different retrieval models. The classical probabilistic model takes feedback documents as positive examples and estimates the model parameters using Bayesian rule (Robertson and Sparck Jones 1977). By combining global and local analysis, Xu and Croft (2000) proposed an effective local analysis algorithm called local context analysis (LCA). In this work, several noun groups are selected from the top ranked documents based on the passage co-occurrence of query terms and introduced into the original query. Despite its popularity in the research community, PRF is likely to deteriorate the retrieval performance when their underlying assumption is violated, i.e., when top-ranked documents are irrelevant to the query.

In contrast to the local strategy, the global strategy is to expand the query description using information from the entire collection or external thesaurus. One global technique is to automatically create a domain-specific thesaurus based on global term-to-term similarities (Qiu and Frei 1993) and use it to expand additional query terms based on their similarities to the query keywords. Other well-known global techniques include latent semantic indexing (Deerwester et al.1990), PhraseFinder (Jing and Croft 1994) and so forth. However, these global analysis techniques must obtain statistics for each pair of terms, which is computationally demanding especially for large text collections. Another type of global analysis methods is to leverage external knowledge sources such as a co-occurrence thesaurus or semantic network. An example of semantic network is WordNet (Fellbaum 1998), an online lexical reference system. Based on these thesauri or semantic networks, we can introduce related terms according to their relationship to the query terms.

However, an external semantic network might not capture the exact term correlation for the specific collections that we are dealing with.

A number of researchers have investigated query expansion techniques for video retrieval. For instance, Chua et al. (2004) evaluated pseudo relevance feedback in the TRECVID'04 dataset, which uses top retrieved documents to obtain a list of additional query keywords and iterate the retrieval process. Their results showed that pseudo relevance feedback can bring a small but not significant improvement over the non-feedback baseline. Kennedy et al. (2005) augmented text retrieval results via two global analysis approaches, i.e., leveraging external knowledge sources of WordNet and Google to enrich the query representation. However, their work did not explicitly report the text retrieval performance after query expansion. To account for high-level semantic concepts, Neo et al. (2006) make use of the WordNet hierarchy and Resnik information-content metric to estimate a heuristic combination weight for semantic concepts. It brings an additional 0.4 MAP improvement over the direct keyword matching approach. A more recent study (Volkmer and Natsev 2006) compared three automatic query expansion techniques including Rocchio-based query expansion, lexical-context based expansion and semantic annotation-based expansion on the TRECVID datasets. Surprisingly, their experiments have underscored the difficulty of automatic query expansion in video collections, because only one out of three approaches can gain higher average precision than the non-expansion baseline. However, they also suggested that combining both the retrieval results without and with query expansion can produce better results than either one of them, especially when the combination is carried out in a query-dependent manner. To summarize, query expansion has shown potential, but further experiments are still necessary to prove that query expansion is able to significantly improve text retrieval performance in video collections.

3.4.1 Experiments

To evaluate the effectiveness of query expansion in the video corpus, we designed and evaluated three types of expansion approaches. The first approach is manual expansion (**Manual**) which asks users to manually introduce additional query words and refine text queries based on development data. The second method is based on a global query expansion strategy (**WordNet**). It passes every query keywords into WordNet and expands a fixed number of synonyms to the original queries. The last method is based on a local query expansion strategy by analyzing the relevance of top retrieved documents (**Local**). The expanded terms are chosen to be the terms with the highest *tf.idf* features in the top 10 retrieved documents.

Table 5 compares all three expansion methods and the baseline method on TRECVID'03-'05 collections. The second column of Table 5 indicates the labels of methods and corresponding expansion parameters, where the parameter following "WordNet" indicates the number of synonyms expanded and the number following "Local" indicates the number of query terms expanded. The message from the experimental results is mixed: manual expansion can considerably boost the retrieval results which shows the usefulness of leveraging additional human knowledge, but the other two types of automatic expansion approaches is not so consistent in producing better performance in terms of average precision, especially when the number of expansion terms grows larger. The inconsistency of the last two approaches can be traced back to the noticeable degradation in precision (even though recall is slightly higher than before). This is because many additional "noisy" terms

Table 5 Comparison of query expansion methods

Data	App.+para.	MAP	P30	P100	R1000	Person	S-Obj	G-Obj	Sports	Others
t05s	Baseline	0.069(+0%)	0.174	0.162	0.340	0.144	0.058	0.016	0.069	0.017
	Manual	**0.103(+49%)**	0.226	0.200	0.383	0.194	0.086	0.013	0.171	0.021
	WordNet 1	0.060(−13%)	0.136	0.138	0.341	0.133	0.052	0.014	0.046	0.011
	WordNet 2	0.059(−14%)	0.135	0.135	0.340	0.133	0.051	0.014	0.046	0.010
	Local 1	0.070(+1%)	0.172	0.163	0.341	0.142	0.057	0.015	0.077	0.017
	Local 2	0.067(−3%)	0.169	0.158	0.341	0.140	0.055	0.014	0.069	0.017
	Local 3	0.065(−6%)	0.165	0.155	0.338	0.136	0.055	0.013	0.062	0.017
t04s	Baseline	0.078(+0%)	0.178	0.105	0.360	0.189	0.000	0.039	0.047	0.041
	Manual	**0.093(+18%)**	0.180	0.118	0.459	0.200	0.000	0.030	0.156	0.044
	WordNet 1	0.052(−33%)	0.103	0.071	0.302	0.104	0.000	0.037	0.027	0.035
	WordNet 2	0.048(−39%)	0.090	0.068	0.296	0.095	0.000	0.035	0.037	0.027
	Local 1	0.071(−9%)	0.165	0.103	0.357	0.169	0.000	0.031	0.047	0.041
	Local 2	0.063(−19%)	0.149	0.102	0.358	0.142	0.000	0.029	0.045	0.041
	Local 3	0.060(−23%)	0.145	0.101	0.358	0.132	0.000	0.027	0.047	0.041
t03s	Baseline	0.150(+0%)	0.184	0.120	0.473	0.372	0.237	0.067	0.033	0.007
	Manual	**0.194(+29%)**	0.227	0.155	0.561	0.404	0.299	0.099	0.135	0.041
	WordNet 1	0.135(−9%)	0.168	0.114	0.490	0.298	0.214	0.081	0.036	0.007
	WordNet 2	0.119(−20%)	0.145	0.098	0.463	0.265	0.185	0.071	0.035	0.006
	Local 1	0.150(+0%)	0.180	0.119	0.473	0.333	0.243	0.066	0.032	0.007
	Local 2	0.143(−4%)	0.180	0.118	0.479	0.351	0.243	0.057	0.032	0.007
	Local 3	0.141(−5%)	0.165	0.116	0.488	0.364	0.235	0.051	0.025	0.007
t05d	Baseline	0.032(+0%)	0.065	0.058	0.344	0.077	0.015	0.006	0.022	0.009
	Manual	**0.056(+75%)**	0.090	0.089	0.407	0.089	0.017	0.034	0.153	0.011
	WordNet 1	0.028(−10%)	0.054	0.052	0.331	0.075	0.010	0.005	0.015	0.008
	WordNet 2	0.028(−12%)	0.053	0.052	0.332	0.075	0.010	0.002	0.015	0.006
t04d	Baseline	0.073(+0%)	0.097	0.075	0.487	0.130	0.000	0.061	0.077	0.036
	Manual	**0.094(+28%)**	0.115	0.084	0.691	0.174	0.000	0.079	0.113	0.031
	WordNet 1	0.056(−23%)	0.074	0.049	0.440	0.065	0.000	0.067	0.069	0.033
	WordNet 2	0.049(−33%)	0.075	0.045	0.417	0.057	0.000	0.064	0.046	0.032
t03d	Baseline	0.092(+0%)	0.077	0.051	0.480	0.185	0.102	0.080	0.048	0.009
	Manual	**0.116(+26%)**	0.092	0.066	0.579	0.199	0.144	0.109	0.084	0.012
	WordNet 1	0.081(−11%)	0.065	0.044	0.508	0.186	0.055	0.079	0.041	0.008
	WordNet 2	0.068(−25%)	0.043	0.036	0.468	0.167	0.030	0.067	0.040	0.008

are introduced into the query after the step of query expansion. Another reason for their inconsistency can be attributed to the subpar retrieval performance in video collections, which prevents automatic query expansion techniques from significantly improving the search results by assuming top-ranked documents are mostly relevant. To further compare their performance w.r.t. each query type, it can be found that the finding-person queries degrade the most with the automatic expansion, which suggests query expansion is not effective to search for person-related shots. Interestingly, the finding-sport queries gain a significant improvement from manual query expansion, indicating the potential of expanding extra sports-related words in query topics. To summarize, the best query expansion approach for a video corpus so far is to expand keywords in a manually controlled manner rather than in an automatic way. Note that, what we attempt to emphasize is the difficulty rather than the failure of using automatic query expansion to improve video retrieval. But in order to apply automatic expansion for video collections, it will be critical to develop more robust automatic expansion techniques than what we used in the experiments.

4 Image retrieval

Content-based image retrieval(CBIR) has been developed for more than a decade (Smeulders et al.2000). Its goal is to search a given image collection for a set of relevant images that are similar to one or more query images. Previous research efforts have led to many successful image retrieval systems such as MARS (Rui et al.1997a), VisualSeek (Smith and Chang 1996c), QBIC (Faloutsos et al.1994), SIMPLicity (Li et al.2000) and so on. In news video retrieval, although CBIR is not so powerful as text retrieval in terms of handling general semantic queries, it is useful in dealing with a number of queries from several specific domains, where information needs are consistent with visual appearances. For instance, CBIR has great success when queries are related to sport events or duplicate commercial shots. Typical CBIR systems are built on a vector space model that represents an image as a set of features. The difference between two images is measured through a similarity function between their feature vectors. They take a few image examples as inputs, convert them into sets of image features, match them with the features of all images in the collection, and retrieve the closest ones to the users. In the rest of this section, we discuss each individual component of image retrieval systems and evaluate them in the context of video collections.

4.1 Image features

Similar to term weights in text retrieval, image features are represented as a vector of real values, which aim to compress high-dimensional image information into a lower dimensional vector space. In the literature, there are mainly three types of (low-level) image features, i.e., color-based features, texture-based features and shape-based features (Antani et al.2002; Smeulders et al.2000; Rui et al. 1997b)[3].

Color-based features have been shown to be the most widely-used features in CBIR systems, because they maintain strong cues that capture human perception in a low dimensional space and they can be generated with less computational effort than other advanced features. Most of them are independent of variations of view and resolution, and thus possess the power to locate the target images robustly. They have also been demonstrated to be the most effective image features in the TRECVID evaluation (Rautiainen et al.2004a; Amir et al.2003; Hauptmann et al.2003a; Foley et al.2005; Cooke et al. 2004; Adcock et al.2004). Many color spaces have been suggested in previous studies such as RGB, YUV, HSV, HVC, L*u*v*, L*a*b*, and the Munsell space (Del Bimbo 2001). The

[3] There are many other approaches to produce image features for content-based image retrieval, but it is not our focus to provide an exhaustive list for all of them. A more complete survey of CBIR can be found at Antani et al. (2002) and Smeulders et al. (2000).

simplest representation of color-based features are color histograms. Each component in color histograms is the percentage of pixels that are most similar to the represented color in the underlying color space (Faloutsos et al. 1994; Smith and Chang 1996c). Another type of color-based image features are called color moments, which only compute the first two or three central moments of color distributions (Stricker and Orengo 1995; Smith and Chang 1996b). They aim to create a compact and effective representation in image retrieval. Huang et al. (1997) proposed the use of color correlograms. A color correlogram expresses the spatial correlation of pairs of colors with distance information, thus making it robust against the change of viewpoint, zoom, and etc.

Texture-based features aim to capture the visual characteristics of homogeneous regions which do not come from a single color or intensity (Smith and Chang 1996a). These regions may have unique visual patterns or spatial pixel arrangements, which gray level or color features in a region may not sufficiently describe. The process of extracting texture-based features often begins with passing images into a number of Gabor or Haar wavelet filters (Lee 1996; Amir et al.2003). The feature vector can then be either constructed by concatenating central moments from multiple scales and orientations into a long vector (Manjunath and Ma 1996; Ngo et al. 2001) or extracting statistics from image distributions directly (Puzicha et al.1997; Thyagarajan et al.1996). In the literature, there are a few review papers that aim to investigate the effectiveness of texture features. For instance, Ohanian and Dubes (1992) compared four types of texture representations and observed that the co-occurrence matrix representation work the best in their test collections. Ma and Manjunath (1995) evaluated the wavelet texture features for image annotation, which includes orthogonal/bi-orthogonal wavelet transform, tree-structured wavelet transform and Gabor wavelet transform. They concluded that Gabor wavelet representation was the best among all the tested features.

To capture information from object shapes, a huge variety of shape-based features have been proposed and evaluated (Zahn and Roskies 1972; Chuang and Kuo 1996; Li and Ma 1994; Mehtre et al.1997). Shape features can be generated either from boundary-based approaches that use only the outer contour of shape segments, or from region-based approaches that considers the entire shape regions (Rui et al.1996). One of the simplest approaches to extract shape features is to detect visible edges in query images and then match their edge distribution or histogram against those of the target images (Marr and Hildreth 1979; Hauptmann et al.2004; Cooke et al. 2004). Another approach to extract shape features is to use implicit polynomials to effectively represent the geometric shape structures (Lei et al.1997), which is robust and stable to general image transformation. Mehtre et al. (1997) presented a comprehensive comparison of shape features for retrieval by evaluating them on a 500-element trademark dataset. Another review paper on shape-based features can be found at (Li and Ma 1994).

It is not always necessary to construct image features by globally extracting features from the entire image. Although global image features are efficient to compute and provide reasonable retrieval capabilities, they are very likely to generate unpredictable false positives due to its concise representations (Rui et al.1997b). In contrast, image features can be extracted from a finer granularity, such as regular image grids/layouts, automatically segmented image blobs and local feature points. In practice, content-based image classification/retrieval based on regional features usually shows better performance than its counterpart using global features, although it might lead to a higher computational cost in the step of feature extraction. In the following discussions, we review some general methods on extracting local image features.

To derive local features from images, a natural idea is to partition the entire image into a set of regular image grids and extract image features (especially color features) from image grids (Amir et al. 2003; Faloutsos et al.1994; Chua et al.1997). For instance, Cooke et al. (2004) used a local color descriptor based on the average color components on an 8 × 8 block partition of images. Hauptmann et al. (2004) studied color layout features on a 5 × 5 regular image grid. A 4 × 4 spatial image grid is used in (Adcock et al.2004). Extended from regular image grids, quad-tree based layout (Lu et al.1994) approaches first split the image into a quad-tree structure and construct color histogram for each tree branch so as to describe its local image content. Although being simple in their intuitions, concepts and implementations, regular-image-grid based approaches could be still too coarse to provide sufficient local information for the retrieval task. Therefore several other image layout representations have been proposed before. For instance, Stricker and Orengo (1995) predefined five partially overlapped regions and extracted the first three color moments from each region, where the advantage of the overlapping regions is their relative insensitivity to small regional transformations. The representation of color tuple histogram is suggested in (Rickman and Stonham 1996), which first builds a codebook to represent every combination of coarsely quantized close hues and then compute a local histogram based on the quantized hues. Color coherent histograms (Pass and Zabih 1999) and color correlograms (Huang et al.1997) are two more examples of advanced representations that take spatial image information into account.

In order to locate specific objects in images, it would be advantageous to extract image features (e.g., color or shape) from segmented image regions. Image segmentation is defined as "a division of the image data into regions in such a way that one region contains the pixels of the silhouette of the given object without anything else." (Smeulders et al.2000). This task is so important in the literature of compute vision that a huge variety of segmentation approaches have been proposed before. Survey of mainly historical interests can be found at (Nevatia 1986; Mitiche and Aggarwal 1985; Pal and Pal 1993). Most segmentation algorithms proceed by automatically clustering the pixels into groups. A number of graphical theoretical clustering approaches have been proposed before (Sarkar and Boyer 1996; Cox et al.1996) due to their ability to deal with any affinity function. For example, normalized cut (NCut) proposed by Shi and Malik (1998, 2000) has been widely applied in visual retrieval, object recognition and motion segmentation. Numerous alternative criteria have also been suggested for segmentation (Perona and Freeman 1998; Cox et al.1996). The use of image segments has been widely studied in the context of content-based video retrieval, such as (Srikanth et al.2005; Zhai et al. 2005; Hauptmann and Christel 2004). Note that, in this case, the requirement of segmentation accuracy is highly dependent on the choice of image features. For the color features, a coarse segmentation should be sufficient, while for the shape features, accurate segmentation is usually desirable. A final comment worth mentioning is that segmenting general objects in broad domains is unlikely to succeed, although there are exceptions for some sophisticated methods in narrow domains (Smeulders et al.2000).

One way to circumvent the brittleness of segmentation but maintain the local information of images is to extract image features from selected salient points (a.k.a. feature points). It aims to concisely summarizes image information into a limited number of salient points, and thus these points should be selected with a high saliency and robustness. In (Carson et al.1997), a mixture of Gaussian models is estimated to model the distribution of salient points. The information of homogeneous regions is captured by means and covariances of the Gaussian components. To improve the feature quality, invariant and salient features of local patches have also been considered (Tuytelaars and van Gool 1999).

In (Schmid and Mohr 1997), salient and invariant transitions are recorded in gray images. To localize all the occurrences of a query object in videos, “Video Google” (Sivic and Zisserman 2003) represents objects by a set of SIFT-based viewpoint invariant descriptors, and thus this recognition technique can work robustly to viewpoint changes, illumination changes and partial occlusion. Chang et al. (2005a) investigated a part-based object representation in TRECVID collections in order to capture spatial relationship and local attributes of salient parts. Zhai et al. (2005) also evaluated the performance of local features from image segments and feature points in video collections.

4.1.1 Experiments

To evaluate image retrieval over the TRECVID corpus, we have extracted three types of low-level features as described above, i.e., color-based features, texture-based features and edge-based features. By default, image features are generated over 5 × 5 regular grids posed on every image. All grid features are concatenated into a longer vector unless stated otherwise. Each dimension of the feature vector is normalized by its own variance. Finally, we compute the harmonic mean of Euclidean distances from each image example to the document keyframes (officially provided by NIST). The details of the feature generation process are shown as follows,

- The color features are computed based on both the HSV color space and the RGB color space. For each space, we extract both a full color histogram with a 5-bin quantization of every color channel and a color moment histogram including the first and second moments.
- The texture features are obtained from the convolution of the image pixels with Gabor wavelet filters. For each filter we computed a histogram which was quantized into 16 bins. Their central and second-order moments are concatenated into a texture feature vector. Two versions of texture features are used in our implementation: one uses 6 filters in a 3 × 3 image grids and the other uses 12 filters for 5 × 5 image grids.
- The edge histogram is summarized from the outputs of a Canny edge detector. It includes a total of 73 bins, where the first 72 bins represent the edge directions quantized at a 5 degree interval and the last bin represents a count of the number of pixels that are not contributing to any edges.

The comparison of various image features are shown in Table 6. As shown in the experiments, image retrieval can only achieve a poor 2% – 3% mean average precision for almost all the collections. Obviously it is not as effective as text retrieval on average. This can be explained by the fact that image features are not as powerful as text features for broadcast news video in terms of capturing semantic meanings. The requirement of searching for shots instead of clips in TRECVID evaluation also limits the applicability of image features. As an exception, image retrieval works better in the latest TREC’05 collection, mainly due to the superior effectiveness of image matching in sport queries.

Among all kinds of color features, color moments in both the HSV/RGB space has the best performance on average, followed by color histograms. This again confirms the effectiveness and explains the popularity of color-based features. Occasionally, edge histograms can provide a comparable performance with color, but its performance is not as consistent across all collections. The texture features, unfortunately, are among the worst due to their inability to capture the semantics in non-texture images. Taking a deeper look at each query type, we found that the best candidates for image retrieval are the queries for

Table 6 Comparison of image features

Data	Feature	Para.	MAP	P30	P100	Person	S-Obj	G-Obj	Sports	Others
t05s	color:hsv	his:5 × 5	**0.039**	0.114	0.078	0.033	0.021	0.003	0.186	0.005
	color:hsv	mom:5 × 5	0.032	0.090	0.057	0.007	0.021	0.003	0.194	0.004
	color:rgb	his:5 × 5	0.037	0.114	0.081	0.019	0.022	0.005	0.202	0.004
	color:rgb	mom:5 × 5	0.034	0.101	0.056	0.004	0.022	0.004	0.211	0.003
	texture	mom:3 × 3	0.005	0.026	0.019	0.001	0.006	0.001	0.028	0.001
	texture	mom:5 × 5	0.003	0.024	0.021	0.000	0.006	0.002	0.009	0.001
	edge	his:5 × 5	0.009	0.042	0.035	0.000	0.014	0.003	0.044	0.003
t04s	color:hsv	his:5 × 5	0.004	0.010	0.007	0.000	0.000	0.002	0.024	0.000
	color:hsv	mom:5 × 5	0.014	0.045	0.028	0.003	0.000	0.001	0.095	0.003
	color:rgb	his:5 × 5	0.008	0.029	0.018	0.004	0.000	0.001	0.055	0.000
	color:rgb	mom:5 × 5	**0.016**	0.048	0.032	0.022	0.000	0.001	0.075	0.001
	texture	mom:3 × 3	0.003	0.010	0.010	0.001	0.000	0.000	0.017	0.001
	texture	mom:5 × 5	0.003	0.006	0.010	0.001	0.000	0.000	0.019	0.000
	edge	his:5 × 5	0.003	0.013	0.011	0.000	0.000	0.002	0.016	0.001
t03s	color:hsv	his:5 × 5	0.026	0.072	0.046	0.000	0.078	0.001	0.113	0.008
	color:hsv	mom:5 × 5	0.035	0.087	0.058	0.000	0.064	0.010	0.221	0.010
	color:rgb	his:5 × 5	0.029	0.043	0.030	0.000	0.114	0.001	0.065	0.002
	color:rgb	mom:5 × 5	**0.049**	0.088	0.060	0.000	0.095	0.010	0.313	0.010
	texture	mom:3 × 3	0.024	0.036	0.026	0.000	0.052	0.002	0.161	0.000
	texture	mom:5 × 5	0.016	0.032	0.027	0.000	0.005	0.004	0.169	0.000
	edge	his:5 × 5	0.028	0.056	0.036	0.001	0.065	0.013	0.124	0.002
t05d	color:hsv	his:5 × 5	0.060	0.144	0.075	0.028	0.082	0.003	0.260	0.009
	color:hsv	mom:5 × 5	0.058	0.125	0.075	0.007	0.072	0.007	0.309	0.004
	color:rgb	his:5 × 5	**0.097**	0.160	0.090	0.028	0.127	0.030	0.451	0.010
	color:rgb	mom:5 × 5	0.076	0.135	0.074	0.017	0.097	0.010	0.385	0.005
	texture	mom:3 × 3	0.029	0.050	0.028	0.004	0.097	0.000	0.051	0.003
	texture	mom:5 × 5	0.026	0.051	0.025	0.003	0.059	0.000	0.100	0.000
	edge	his:5 × 5	0.029	0.075	0.035	0.010	0.038	0.001	0.146	0.001
t04d	color:hsv	his:5 × 5	0.016	0.036	0.020	0.023	0.000	0.007	0.051	0.000
	color:hsv	mom:5 × 5	0.018	0.051	0.023	0.029	0.000	0.007	0.050	0.001
	color:rgb	his:5 × 5	**0.021**	0.057	0.025	0.033	0.000	0.005	0.065	0.002
	color:rgb	mom:5 × 5	0.016	0.050	0.025	0.034	0.000	0.001	0.039	0.005
	texture	mom:3 × 3	0.001	0.004	0.003	0.000	0.000	0.000	0.007	0.000
	texture	mom:5 × 5	0.001	0.011	0.005	0.004	0.000	0.000	0.001	0.000
	edge	his:5 × 5	0.016	0.051	0.024	0.035	0.000	0.000	0.034	0.005
t03d	color:hsv	his:5 × 5	0.031	0.039	0.018	0.006	0.130	0.004	0.027	0.001
	color:hsv	mom:5 × 5	0.034	0.035	0.022	0.006	0.131	0.005	0.053	0.002
	color:rgb	his:5 × 5	**0.038**	0.039	0.021	0.006	0.153	0.008	0.032	0.002
	color:rgb	mom:5 × 5	0.044	0.053	0.026	0.006	0.153	0.012	0.088	0.006
	texture	mom:3 × 3	0.006	0.011	0.005	0.000	0.005	0.000	0.064	0.000
	texture	mom:5 × 5	0.005	0.011	0.004	0.000	0.001	0.002	0.058	0.000
	edge	his:5 × 5	0.029	0.044	0.017	0.000	0.070	0.019	0.109	0.000

Table 6 continued

Data	Feature	Para.	MAP	P30	P100	Person	S-Obj	G-Obj	Sports	Others
t02s	color:hsv	mom:5 × 5	**0.029**	0.057	0.033	0.127	0.012	0.008	0.000	0.015
	texture	mom:5 × 5	0.006	0.021	0.016	0.009	0.003	0.006	0.000	0.007

finding specific objects and sport events. This is reasonable because target specific objects and sport events usually share consistent visual appearance with the given image examples. But image retrieval is not a good idea for the other three types of queries, i.e., person, object and other general queries.

4.2 Distance metric

Image retrieval algorithms usually sort and retrieve relevant images based on a predefined similarity measure (distance metric) between query examples and indexed images. Previous studies show that the choice of similarity measure is critical to image retrieval performance (Antani et al.2002). Thus, a large number of distance metrics have been proposed and tested in the literature. In the following discussion, we discuss several common distance metrics with the assumption that only one query image is available.

Two widely used distance metrics, i.e., Euclidean distance (L_2 distance) and absolute distance (L_1 distance), are both special cases of the L_M metric or the Minkowski distance metric (Smeulders et al.2000). An extension of the Euclidean distance is the Mahalanobis distance, where the inverse of a covariance matrix C is plugged into the quadratic function and associated different weights to each feature dimension (Antani et al.2002). If the underlying features are computed in the form of histograms, retrieval systems usually adopt some distance metrics that capture the difference between two probability distributions. For example, we can compute simple absolute difference of feature histograms. As a more reliable distance metric w.r.t. histogram features, color histogram intersection was proposed for image retrieval (Swain and Ballard 1991), where a value close to 1 indicates high similarity. The χ^2 distance for comparing two histograms was proposed by Nagasaka and Tanaka (1992) where a low value indicates a good match. Its underlying idea is to find the images with histogram distributions least independent to the query examples. Stricker (1994) has studied the discrimination ability of histogram-based indexing methods. He concluded in his work that the histogram-based technique would only work effectively when the histograms are sparse. Beyond using fixed distance metrics, numerous relevance feedback and manifold learning approaches (Rui et al.1997a; He et al. 2004, 2002; Su et al.2001) have also been proposed to learn distance metrics adaptively based on information from user feedback.

Most image retrieval algorithms simply consider dealing with one query example at a time. But since it is not impossible that users could simultaneously provide multiple image examples to the retrieval systems, we might need to come up with some approaches to aggregate all of the distance metrics from each query image to be a final ranked list. The common approach is to measure image similarities from individual query images and then fuse the similarity measures into a single output via certain kinds of fusion methods. We will consider five types of common aggregation functions in our following experiments, i.e., maximum, minimum, harmonic mean, average (arithmetic mean), and product (geometric mean) (Amir et al. 2003; Hauptmann et al.2003a). Several advanced multi-query-example retrieval approaches have also been proposed before. McDonald and Smeaton (2005) studied the effect of various combination strategies for merging multiple visual examples. Jin and Hauptmann (2002) proposed a probabilistic image retrieval model by computing the conditional probability of generating the target image given multiple query images. Westerveld and de Vries (2004) developed a document generation model to handle multi-example queries, which capture all the information available in the query examples with a limited number of Gaussian components. Natsev and Smith (2003) considered three types of criteria to automatically select the most effective image examples for retrieval, i.e., KMEANS which uses the mean of image clusters as queries, MINDIST which finds the most distinct positive examples greedily, and SUMDIST which provide a compromise between KMEANS and MINDIST criteria.

4.2.1 Experiments

We compared three types of distance metrics including the L_2, L_1 and χ^2 metrics in Table 7. The underlying image features are chosen to be color moments on the HSV color space. Each dimension is normalized by its own variance. We observe that using the L_1 distance metric usually work slightly better than using the L_2 and χ^2 distance, but their differences are not statistically significant. Therefore, it is not conclusive yet to judge which metric is the best choice for image retrieval. But being robust to outliers and

Table 7 Comparison of image distance metrics

Data	Dist.	MAP	P30	P100	Person	S-Obj	G-Obj	Sports	Others
t05s	L_2	0.025	0.076	0.048	0.010	0.021	0.002	0.130	0.003
	L_1	**0.031**	0.089	0.060	0.008	0.022	0.003	0.183	0.004
	χ^2	0.025	0.085	0.055	0.007	0.014	0.003	0.146	0.003
t04s	L_2	0.014	0.045	0.028	0.003	0.000	0.001	0.095	0.003
	L_1	**0.017**	0.046	0.032	0.007	0.000	0.002	0.107	0.002
	χ^2	0.015	0.043	0.031	0.002	0.000	0.001	0.097	0.002
t03s	L_2	0.035	0.087	0.058	0.000	0.064	0.010	0.221	0.010
	L_1	**0.039**	0.085	0.055	0.000	0.079	0.007	0.238	0.009
	χ^2	0.037	0.079	0.056	0.001	0.076	0.010	0.210	0.010
t05d	L_2	0.044	0.096	0.061	0.006	0.066	0.003	0.216	0.004
	L_1	0.055	0.117	0.073	0.007	0.077	0.005	0.289	0.003
	χ^2	**0.057**	0.124	0.064	0.009	0.094	0.006	0.264	0.003
t04d	L_2	0.018	0.051	0.023	0.029	0.000	0.007	0.050	0.001
	L_1	**0.021**	0.057	0.025	0.035	0.000	0.006	0.058	0.005
	χ^2	0.018	0.047	0.023	0.026	0.000	0.007	0.057	0.001
t03d	L_2	0.034	0.035	0.022	0.006	0.131	0.005	0.053	0.002
	L_1	**0.040**	0.037	0.026	0.006	0.154	0.009	0.055	0.002
	χ^2	0.028	0.032	0.019	0.006	0.092	0.009	0.055	0.002
t02s	L_2	**0.029**	0.057	0.034	0.129	0.012	0.008	0.000	0.015
	L_1	**0.029**	0.052	0.036	0.126	0.016	0.009	0.000	0.011
	χ^2	**0.029**	0.052	0.031	0.136	0.008	0.008	0.000	0.018

efficient to compute (Leroy and Rousseeuw 1987), the L_1 distance appears to be one of the most effective metrics in practice.

Table 8 compares several fusion functions that are used to merge the retrieval outputs from multiple query images. The distance metric is set to be the L_1 distance. It can be observed that the harmonic mean and maximum functions outperform the other fusion functions in terms of mean average precision. Their superior performance can be attributed to a nice property: they tend to give a higher rank to images that are very close to one of the

Table 8 Comparison of query example fusion strategies

Data	Merge	MAP	P30	P100	Person	S-Obj	G-Obj	Sports	Others
t05s	Harmonic	0.032	0.090	0.057	0.007	0.021	0.003	0.194	0.004
	Maximum	**0.035**	0.110	0.064	0.019	0.019	0.003	0.194	0.004
	Minimum	0.002	0.017	0.015	0.004	0.000	0.001	0.000	0.000
	Average	0.027	0.079	0.049	0.002	0.009	0.003	0.190	0.003
	Product	0.031	0.085	0.054	0.003	0.024	0.003	0.192	0.004
t04s	Harmonic	0.020	0.048	0.037	0.007	0.000	0.001	0.129	0.002
	Maximum	**0.022**	0.055	0.039	0.004	0.000	0.001	0.152	0.004
	Minimum	0.001	0.003	0.004	0.000	0.000	0.001	0.000	0.002
	Average	0.013	0.033	0.027	0.006	0.000	0.001	0.083	0.001
	Product	0.017	0.042	0.030	0.007	0.000	0.001	0.110	0.002
t03s	Harmonic	**0.044**	0.087	0.057	0.000	0.100	0.008	0.247	0.009
	Maximum	0.037	0.071	0.052	0.000	0.102	0.006	0.158	0.007
	Minimum	0.002	0.005	0.003	0.000	0.000	0.005	0.000	0.002
	Average	0.034	0.076	0.053	0.000	0.057	0.003	0.252	0.009
	Product	0.038	0.080	0.057	0.000	0.073	0.004	0.251	0.009
t05d	Harmonic	0.058	0.125	0.075	0.007	0.072	0.007	0.309	0.004
	Maximum	**0.084**	0.169	0.084	0.034	0.125	0.025	0.328	0.016
	Minimum	0.000	0.000	0.001	0.000	0.000	0.000	0.000	0.000
	Average	0.049	0.121	0.064	0.003	0.051	0.005	0.290	0.003
	Product	0.053	0.111	0.066	0.003	0.063	0.005	0.299	0.003
t04d	Harmonic	0.022	0.058	0.027	0.033	0.000	0.007	0.066	0.005
	Maximum	**0.026**	0.065	0.029	0.039	0.000	0.008	0.074	0.005
	Minimum	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	Average	0.008	0.021	0.015	0.004	0.000	0.000	0.043	0.000
	Product	0.017	0.046	0.023	0.026	0.000	0.001	0.059	0.001
t03d	Harmonic	**0.040**	0.043	0.026	0.006	0.154	0.009	0.061	0.002
	Maximum	0.039	0.036	0.019	0.006	0.162	0.009	0.027	0.002
	Minimum	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	Average	0.028	0.033	0.023	0.002	0.109	0.003	0.053	0.002
	Product	0.035	0.039	0.025	0.006	0.135	0.005	0.054	0.002
t02s	Harmonic	0.029	0.051	0.037	0.120	0.019	0.009	0.000	0.010
	Maximum	0.025	0.044	0.031	0.111	0.010	0.008	0.000	0.007
	Minimum	0.002	0.003	0.002	0.002	0.001	0.001	0.000	0.001
	Average	0.024	0.035	0.028	0.129	0.003	0.004	0.000	0.009
	Product	**0.030**	0.056	0.035	0.127	0.020	0.008	0.000	0.010

query images, even if they are far away from others. In some sense, the harmonic mean and maximum functions are similar to a noisy "logical-OR" operator on a set of Boolean similarity predictions. This property is extremely important, especially when relevant images only share similar visual patterns with *one* of the query images rather than *all* of them.

5 Beyond text and image retrieval: concept-based retrieval and multimodal combination

In this section, we briefly discuss two additional topics beyond text and image retrieval for video collections, i.e., an emerging video retrieval approach based on high-level semantic concepts (termed concept-based retrieval), and the approaches to combine retrieval outputs from multiple modalities.

5.1 Concept-based retrieval

The image/video analysis community has long struggled in bridging the semantic gap from low-level features to high-level semantic content. To overcome this gap, recent years have seen the emergence of a new retrieval approach, called concept-based retrieval, which aims to design and utilize a set of intermediate semantic concepts (Naphade and Smith 2004) to describe frequent visual content in video collections and improve the retrieval performance. These concepts cover a wide range of topics (Chang et al.2005b) such as those related to people (face, anchor, etc), acoustics (speech, music, significant pause), objects (image blobs, buildings, graphics), location (outdoors/indoors, cityscape, landscape, studio setting), genre (weather, financial, sports) and production (camera motion, blank frames). In the following, we briefly describe how to detect high-level semantic concepts and how to use these concepts to assist video retrieval.

5.1.1 Semantic concept detection

The task of semantic concept detection has been investigated by many studies (Barnard et al.2002; Naphade et al.1998; Lin et al.2003; Jeon et al.2003; Wu et al.2004; Szummer and Picard 2002). Their successes have demonstrated that a large number of high-level semantic concepts are able to be inferred from low-level multi-modal features. Typically, the first step of developing semantic concept detection systems is to define a meaningful and manageable list of semantic concepts based on human prior knowledge. For each individual concept, we should manually collect its ground truth on a development video collection. For example, the common annotation forum in TRECVID'03 has successfully annotated 831 semantic concepts on a 65 h development video collection (Lin et al.2003).

Most previous work approached concept detection as a supervised learning problem that attempts to discriminate positive and negative annotated examples through automatically extracted low-level features. As the first step, a variety of low-level features are extracted from several modalities, e.g., text, audio, motion and visual modality. For each concept, separate uni-modal classifiers are built using the corresponding labeled data and low-level features. One of the most common learning algorithms is called support vector machines (SVMs) (Burges 1998), which have been proposed with sound theoretical justifications so

Table 9 Average precision of using SVMs to detect 22 frequent semantic concepts. The model is learned from the TREC'05 development data with color moment features. The column "positive" indicates the number of positive examples of each concept out of 55932 training shots

Concept	Avg Prec	Positive	Concept	Avg Prec	Positive
Person	0.8531	31161	Road	0.2481	2665
Face	0.7752	17337	Microphone	0.1947	2659
Outdoor	0.7114	15290	Interview	0.3019	2619
Studio	0.7541	4743	InterviewSeq	0.5237	2523
Building	0.3048	4177	Car	0.3151	2492
Female	0.2632	3887	Meeting	0.1708	2262
Walking	0.1635	3828	Anchor-Studio	0.8247	2392
Urban	0.1127	3586	Artficial-Text	0.6783	2373
Leader	0.1822	3033	Trees	0.2522	2152
Politicians	0.2782	2850	Sports	0.4481	1249
Asian-people	0.4247	2776	Maps	0.4816	610

as to provide good generalization performance. Apart from SVMs, there are a large variety of other classifiers (Naphade and Smith 2004) that have been investigated, including Gaussian mixture models (GMM), hidden Markov models (HMM), k Nearest Neighbor (kNN), logistic regression, Adaboost and so on. To illustrate, Table 9 shows the average precision of detecting several frequently occurring semantic concepts for TRECVID'05 development data using SVMs (Yan 2006).

To further refine the detection results, it is beneficial to combine prediction outputs from multiple modalities that provide complementary information with each other. Generally speaking, there are two families of multi-modal fusion approaches, i.e., early fusion and late fusion. The early fusion method begins with merging multi-modal features into a longer feature vector and takes as the input of learning algorithms. In contrast, the late fusion method directly fuses the detection outputs from multiple uni-modal classifiers. Both fusion methods have their own strengths and weaknesses (Snoek et al.2005), but late fusion appears to be more popular and more extensively studied than early fusion in the literature. Finally, since the detection results of semantic concepts are not related to any query topic, they can be indexed offline without consuming any online computation resources. Such detection approaches have been applied in most existing video semantic concept extraction systems (Hauptmann et al.2003a; Amir et al. 2003).

Due to the space limit, we refer to a survey written by Naphade and Smith (2004) and Chapter 3 in our previous work (Yan 2006) for more details. Note that, in order to utilize additional domain knowledge for some widely applicable concepts such as faces, cars and sport events, researchers also developed numerous domain-specific recognition approaches in a case-by-case basis. But since the discussions on domain-specific techniques are outside the scope of this article, we will skip this topic.

5.2 Video retrieval with semantic concepts

To illustrate how semantic concepts can be used in video retrieval, we discuss four most common types of concept-based retrieval methods. The simplest approach is to match the name of each concept with query terms. If a concept is found to be relevant, its detection outputs can be used to refine the retrieval results. For example, the concept “building“ will be helpful for retrieving the query of “finding the scenes containing buildings in New York City”. This method is intuitive to understand and simple to implement. However, it is unrealistic to expect a general user to explicitly indicate all related concepts in his query text. For example, the concept of “outdoor“ could be useful for the query of “finding people on the beach”, but it does not show up in the query text directly.

To extend the power of simple query matching, we can follow the idea of global query analysis in text retrieval, which attempts to enrich query descriptions from external knowledge sources, such as a semantic network (ontology) organized to provide semantic relations between keywords, e.g., WordNet (Fellbaum 1998). These approaches have shown promising retrieval results (Volkmer and Natsev 2006; Neo et al. 2006) by introducing extra concepts from external knowledge sources. However, they are also likely to bring in other noisy concepts, and thus it might lead to unexpected deterioration of search results. Moreover, even when the subset of relevant concepts are perfectly detected, it remains a challenge to derive a good strategy to combine semantic concepts with other text/image retrieval results.

As an alternative, we can leverage semantic concepts by learning the combination strategies from a pre-collected training collection, e.g., learning query-independent combination models (Amir et al. 2003) and query-class dependent combination models (Yan et al.2004). These approaches can automatically determine concept weights and handle hidden semantic concepts without any difficulties. However, since these learning approaches can only capture the general patterns that distinguish relevant and irrelevant training documents, their power is usually limited by the amount of available training data.

Finally, we can also consider local analysis approaches that adaptively leverage semantic concepts on a per query basis. The essence of local strategies is to utilize initial retrieved documents to select expanded discriminative query concepts to improve the retrieval performance. For example, we proposed a retrieval approach called probabilistic local context analysis (pLCA) (Yan 2006), which can automatically leverage useful high-level semantic concepts based on initial retrieval output. However, the success of these approaches usually relies on reasonably accurate initial search results. If initial retrieval performance is unsatisfactory, it is possible for local analysis approaches to degrade the retrieval results.

To summarize, all four types of approaches have proved to be successful in utilizing high-level semantic concepts for video retrieval, although they all come with their own limitations. Moreover, the applicabilities of these methods are not mutually exclusive. Instead, a composite strategy can usually produce better results than any single approach. How to automatically determine the best strategy or strategies to incorporate high-level concepts into video retrieval is an interesting direction for future exploration.

5.3 Multimodal combination strategies

Designing combination approaches for multiple information sources such as text retrieval and image retrieval is of great importance to develop effective video retrieval systems (Chang et al.2005b). Westerveld et al. (2003) demonstrated how combining different models/modalities can affect the performance of video retrieval. They adopt a generative model inspired by language modeling approach and a probabilistic approach for image retrieval to rank the video shots. Final results are obtained by sorting the joint probabilities of both modalities. The video retrieval system proposed by Amir et al. (2003) applied a query-dependent combination model that the weights are decided based on user experience and a query-independent linear combination model to merge the text/image retrieval systems, where the per-modality weights are chosen to maximize mean average precision on development data. Gaughan et al. (2003) ranked video clips based on the summation of semantic feature and automatic speech retrieval outputs, where the influence of speech retrieval is at four times that of any other features. The QBIC system (Faloutsos et al.1994) combines scores from different image techniques using linear combination.

In the field of text retrieval, Shaw and Fox (1994) proposed a number of combination techniques named COMB{MIN, MAX, ANZ, MED, SUM, MNZ}. The best performing strategies are among COMBSUM (equivalent to averaging), i.e., taking the sum of scores, and COMBMNZ (equivalent to weighted averaging), i.e., multiplying this sum by the number of inputs that have non-zero scores. Vogt and Cottrell (1999) experimented with the weighted linearly combination to merge multiple relevance scores. The learned weights are independent of the queries. Aslam and Montague (2001) proposed a probabilistic model based on Bayes inference using rank information instead of scores, which achieves good results but requires extensive training efforts. They also studied a rank-aggregation approach called the (weighted) Borda-fuse stemmed from the Social Choice Theory.

Query-based combination approaches (Yan et al.2004; Chua et al.2004) have been recently proposed as a viable alternative for the query independent combination strategies, which begins with classifying queries into predefined query classes and then applies the corresponding combination weights to combine knowledge sources. Experimental evaluations have demonstrated the effectiveness of this idea, which have been applied in the best-performed systems of TRECVID manual retrieval task (Smeaton and Over 2003) from the year of 2003. Also, the validity of using query-class dependent weights has been confirmed by many follow-on studies (Chua et al.2005; Huurnink 2005; Yuan et al.2005; Kennedy et al.2005). For example, Huurnink (2005) suggested it is helpful to categorize the queries into general/special queries and simple/complex queries for combination. Yuan et al. (2005) classified the query space into person and non-person queries in their multimedia retrieval system. To improve upon the manually defined query classes, Kennedy et al. (2005) recently proposed a data-driven learning approach to automatically discover the query-class-dependent weights from training data by means of grouping the queries in a joint performance and semantic space via statistical clustering techniques such as hierarchical clustering and k-means. More recent work (Yan and Hauptmann 2006) unified query class categorization and combination weight optimization in a single probabilistic framework by treating query classes as latent variables.

5.3.1 Experiments

In this section, we investigate two types of combination strategies, i.e., query independent and query-class dependent strategies, to combine text retrieval and image retrieval results. In more detail, let s be the overall retrieval score, s_t be the text retrieval score, s_i be the image retrieval scores and λ be the combination factor, then we decide the overall retrieval score by the following formula $s = s_t + \lambda s_i$, where λ varies from 0 to 1 in the experiments. McDonald and Smeaton (2005) have shown that the weighted sum scheme is among the most effective approaches for text/image retrieval combination. Before the combination process, the confidence scores from different modalities/models usually need to be normalized into a uniform output space. Typical normalization schemes include rank normalization (Yan et al.2004), range normalization (Amir et al. 2003) and logistic normalization (Platt 1999). In this study, we choose rank normalization to calibrate the retrieval outputs.

Figure 5(left) shows the learning curves for mean average precision (MAP) with different combination factors based on a query-independent combination strategy, namely, the combination is blindly applied for all queries on the TRECVID'03-'05 data collections. In lieu of being consistently improved with larger combination factors λ, the retrieval performance after combination usually drops when λ is larger than a small value around 0.1 – 0.2. For the collections of t04s and t04d, the degradation is pretty noticeable where the average precision at λ = 1 is even worse than using text retrieval alone. This indicates the query-independent combination strategy needs to be further refined in order to provide a consistent performance improvement. In contrast, Fig. 5(right) shows the learning curve based on a simple query-class dependent combination strategy, namely, the combination is only applied on the queries that belong to the classes of finding specific objects and persons. As can be seen, this simple query-class combination strategy consistently achieves higher average precision with a higher combination factor on image retrieval. The overall improvement is around a reasonable 2% (which is statistically significant) given the relatively inferior performance of image retrieval techniques. These series of experiments demonstrate the effectiveness and consistency of handling the combination method in a query-dependent way.

To further analyze how incorporating high-level semantic concepts will affect the combination results, we generated 14 high-level semantic concepts on each video document in addition, i.e., *face, anchor, commercial, studio, graphics, weather, sports, outdoor, person, crowd, road, car, building and motion*. The details on the concept generation process can be found in (Hauptmann et al. 2004). Table 10 compares various retrieval sources and fusion strategies in terms of mean average precision (MAP) and precision/recall at 30 and 100 shots (P30 and P100). We compare two combination approaches, i.e., query-independent combination (**QInd**), and query-class based combination (**QClass**) on multiple retrieval sources including text retrieval, image retrieval and semantic concepts. The details of the combination algorithms can be found at (Yan 2006). We learned the combination parameters from development sets, and applied them to the corresponding search sets. We can observe that semantic concepts can bring a significant improvement over the fusion results based on text and image retrieval. Moreover, we find that the improvement of QInd is not as stable as that of QClass, especially when semantic concepts

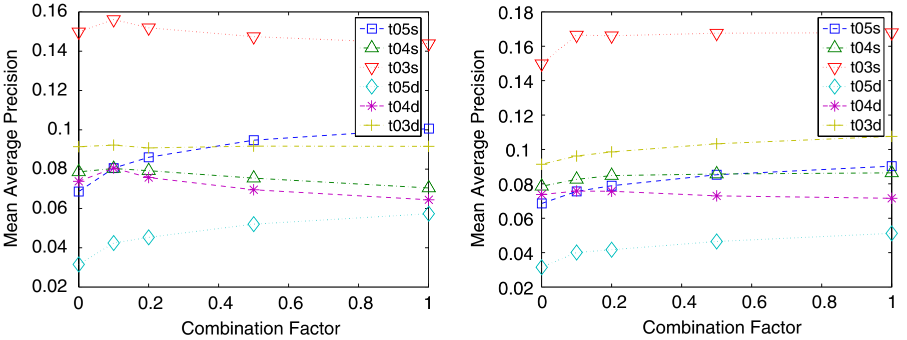


Fig. 5 Comparison of mean average precision for (left) query independent combination and (right) query-class dependent combination against the combination factor λ

Table 10 Comparison of two combination approaches, i.e., query-independent combination (QInd) and query-class based combination (QClass), on multiple retrieval sources including text retrieval (T), image retrieval (I) and semantic concepts (C)

Data	Source	Fusion	MAP	P30	P100	Person	SObj	GObj	Sport	Other
t03s	Text	N/A	0.146(+0%)	0.171	0.118	0.371	0.230	0.068	0.031	0.007
	T+I	QInd	0.166(+14%)	0.183	0.126	0.371	0.309	0.068	0.093	0.007
	T+I	QClass	0.176(+20%)	0.201	0.127	0.404	0.301	0.080	0.091	0.010
	T+I+C	QInd	0.150(+2%)	0.212	0.136	0.251	0.293	0.095	0.101	0.012
	T+I+C	QClass	0.200(+37%)	0.236	0.137	0.407	0.336	0.088	0.106	0.015
t04s	Text	N/A	0.078(+0%)	0.178	0.107	0.188	0.012	0.033	0.046	0.044
	T+I	QInd	0.080(+1%)	0.162	0.108	0.174	0.021	0.028	0.093	0.048
	T+I	QClass	0.084(+7%)	0.191	0.110	0.188	0.024	0.033	0.095	0.044
	T+I+C	QInd	0.079(+0%)	0.177	0.116	0.144	0.063	0.034	0.108	0.051
	T+I+C	QClass	0.094(+20%)	0.199	0.125	0.194	0.080	0.046	0.108	0.045
t05s	Text	N/A	0.073(+0%)	0.207	0.175	0.141	0.015	0.097	0.075	0.016
	T+I	QInd	0.094(+28%)	0.232	0.178	0.172	0.023	0.104	0.151	0.015
	T+I	QClass	0.085(+16%)	0.225	0.181	0.141	0.024	0.097	0.161	0.016
	T+I+C	QInd	0.105(+44%)	0.268	0.205	0.164	0.029	0.090	0.271	0.017
	T+I+C	QClass	0.116(+58%)	0.292	0.211	0.173	0.031	0.100	0.322	0.017

are taken into account. For instance, in the collection of t03s and t03s, learning query-independent weights only produces a poor performance close to the text baseline. In contrast, QClass is almost always superior to QInd in terms of all the measures. The margin between QClass and QInd is quite significant for some collections such as t03s and t05s. This again shows that query-class dependent combination can produce better and more consistent retrieval outputs than query-independent combination, especially when a large number of retrieval sources is available.

6 Conclusion

This paper described and compared state-of-the-art text retrieval and image retrieval approaches in the context of broadcast news video. Numerous components of text/image retrieval have been discussed in detail, including retrieval models, text sources, expansion window size, query expansion, images features and similarity measures. To provide a more complete coverage for video retrieval, we also briefly discuss and experiment with an emerging approach called concept-based video retrieval, and the strategies to combine multiple retrieval outputs. We evaluated the retrieval performance of these components based on multiple TRECVID video collections.

Our experiments have confirmed the following conclusions for video collections: in text retrieval, Okapi models usually provide better performance than vector space models, as is similar to previous experiments in text collections. However, unlike text collections, text retrieval in video corpora is relatively insensitive to the choice of document length normalization schemes. Among five predefined query types, text retrieval is the most effective in queries for finding persons and specific objects. Among all available text sources, closed caption has the best performance, but speech transcript also achieves comparable results in term of average precision, even with a word error rate around 20%. VOCR is shown to be useful in person-finding queries but not in others. Putting all text sources together is often superior to using any single source except in some rare cases. Expanding the text retrieval results to neighbor shots is an important strategy to mitigate the issue of timing misalignment between video clips and relevant text keywords. But for news video retrieval, it is beneficial to limit the expansion size inside the story boundary. Manual query expansion with careful keyword selection can considerably improve text retrieval performance, especially for the queries related to sports events. But on the other hand, automatic query expansion based on WordNet or local feedback can degrade the retrieval performance if it is not handled properly.

In image retrieval, color-based features (especially color moment) are among the best, given their high effectiveness and low computational cost. Edge histogram can occasionally provide a comparable performance with color-based features, but its performance is not as consistent. For each query type, image retrieval is particularly useful for specific object and sport queries, of which the information need can be captured by the visual appearance of a limited number of image examples. But it produces relatively poor performance on the other query types. Being robust to outliers and efficient to compute, the L_1 distance is shown to be one of the most effective distance metrics in practice. To combine multiple query images, using the harmonic mean and maximum function outperforms the other fusion functions in terms of mean average precision. Finally, concept-based retrieval offers another useful way to tackle the video retrieval problem. Combining the outputs of multiple modalities can consistently improve the retrieval performance over any single modality. Meanwhile, it is more robust and effective if the combination is done in a query-dependent way.

It is worthwhile to point out that the focus of this paper is to review text/image retrieval approaches for video collections, and hence it leaves out discussions on other aspects such as human-computer interface, data indexing, storage and so forth. Moreover, this paper mainly investigates video retrieval on broadcast news, instead of other genres such as movie, sport broadcast and medical video. We believe additional studies on the remaining topics will give us a more comprehensive understanding for video retrieval.

Although a huge body of text/image retrieval approaches has been investigated in the literature, the retrieval performance for video collections still have a considerable room to improve. In many ways our research has only just begun. In retrospect, perhaps we have only done the obvious things until this point, but the fundamental “semantic gap” and imperfect text information have still impeded the effectiveness of general text/image retrieval algorithms in practice. We expect the advance of video retrieval technologies in the future will bring more viable solutions to bridge the “semantic gap“ and make video collections more broadly accessible to general users.

Acknowledgements This material is based upon work funded in part by the US Government. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the US Government. The authors thank anonymous reviewers for providing helpful comments to improve this article.

References

Adcock, J., Girgensohn, A., Cooper, M., Liu, T., Wilcox, L., & Rieffel, E. (2004). FXPAL Experiments for TRECVID 2004. In *Proceedings of NIST TREC Video Retrieval Evaluation*, Gaithersburg, MD.

Amir, A., Hsu, W., Iyengar, G., Lin, C. Y., Naphade, M., Natsev, A., Neti, C., Nock, H. J., Smith, J. R., Tseng, B. L., Wu, Y., & Zhang, D. (2003). IBM research TRECVID-2003 video retrieval system. In *Proceedings of NIST TREC Video Retrieval Evaluation*, Gaithersburg, MD.

Antani, S., Kasturi, R., & Jain, R. (2002). A survey on the use of pattern recognition methods for abstraction, indexing and retrieval of images and video. *Pattern Recognition, 4*, 945–65.

Aslam, J. A., & Montague, M. (2001). Models for metasearch. In *Proceedings of the 24th ACM SIGIR conference on Research and development in information retrieva* (pp. 276–284). New Orleans, Louisiana.

Baeza-Yates, R., & Ribeiro-Neto, B. (1999). Modern information retrieval. Reading, MA: Addison Wesley.

Barnard, K., Duygulu, P., Forsyth, D., de Freitas, N., Blei, D., & Jordan, M. (2002). Matching words and pictures. *Journal of Machine Learning Research, 3*, 1107–1135.

Buckley, C., & Walz, J. (1999). SMART in TREC 8. In *Proceedings of the 8th Text REtrieval Conference (TREC)*, Gaithersburg, MD.

Burges, C. (1998). A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery, 2*(2), 955–974.

Carson, C., Belongie, S., Greenspan, H., & Malik, J. (1997). Region-based image querying. In *Proceedings of the 1997 Workshop on Content-Based Access of Image and Video Libraries (CBAIVL '97)* (pp. 42–49). San Juan, Puerto Rico.

Chang, S. F., Hsu, W., Kennedy, L., Xie, L., Yanagawa, A., Zavesky, E., & Zhang, D. (2005a). Columbia university TRECVID-2005 video search and high-level feature extraction. In *Proceedings of NIST TREC Video Retrieval Evaluation*, Gaithersburg, MD.

Chang, S. F., Manmatha, R., & Chua, T. S. (2005b). Combining text and audio-visual features in video indexing. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Philadelphia, PA.

Chen, D., & Odobez, J. M. (2005). Video text recognition using sequential monte carlo and error voting methods. *Pattern Recognition Letter, 26*(9), 1386–1403.

Christel, M., & Hauptmann, A. G. (2005). The use and utility of high-level semantic features. In *Proceedings of International Conference on Image and Video Retrieval (CIVR)*, Singapore.

Christel, M., & Martin, D. (1998). Information visualization within a digital video library. *Journal of Intelligent Information Systems, 11*(3), 235–257.

Chua, T. S., Tan, K. L., & Ooi, B. C. (1997). Fast signature-based color-spatial image retrieval. In *Proceedings of the 1997 International Conference on Multimedia Computing and Systems (ICMCS'97), IEEE Computer Society* (pp. 362–369). Washington, DC, USA .

Chua, T. S., Neo, S. Y., Li, K., Wang, G. H., Shi, R., Zhao, M., Xu, H., Gao, S., & Nwe, T. L. (2004). TRECVID 2004 search and feature extraction task by NUS PRIS. In *Proceedings of NIST TREC Video Retrieval Evaluation*, Gaithersburg, MD.

Chua, T. S., Neo, S. Y., Goh, H. K., Zhao, M., Xiaom Y., & Wang, G. (2005). TRECVID 2005 by NUS PRIS. In *Proceedings of NIST TREC Video Retrieval Evaluation*, Gaithersburg, MD.

Chuang, C. H., & Kuo, C. C. (1996). Wavelet descriptor of planar curves: Theory and applications. *IEEE Transactions on Image Processing, 5*(1), 56–70.

Cooke, E., Ferguson, P., Gaughan, G., Gurrin, C., Jones, G., Borgue, H. L., Lee, H., Marlow, S., McDonald, K., McHugh, M., Murphy, N., O'Connor, N., O'Hare, N., Rothwell, S., Smeaton, A., & Wilkins, P. (2004). TRECVID 2004 experiments in Dublin City University. In *Proceedings of NIST TREC Video Retrieval Evaluation*, Gaithersburg, MD.

Cox, I. J., Rao, S. B., & Zhong, Y. (1996). Ratio regions: A technique for image segmentation. In *Proceedings of International Conference on Pattern Recognition, vol 2* (pp. 557–564).

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science 41*(6).

Del Bimbo, A. (2001). *Visual Information Retrieval*. Morgan Kaufmann Publishers.

Faloutsos, C., Barber, R., Flickner, M., Hafner, J., Niblack, W., Petkovic, D., & Equitz, W. (1994) Efficient and effective querying by image content. *Journal of Intelligent Information Systems, 3*(3/4), 231–262.

Fellbaum, C. (1998). *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.

Foley, C., Gurrin. C,, Jones, G. Lee, H., McGivney, S., O'Connor, N. E., Sav, S., Smeaton, A. F., & Wilkins, P. (2005). TRECVID 2005 experiments in Dublin City University. In *Proceedings of NIST TREC Video Retrieval Evaluation*, MD: Gaithersburg.

Gaughan, G., Smeaton, A. F., Gurrin, C., Lee, H., & McDonald, K. (2003). Design, implementation and testing of an interactive video retrieval system. In *Proceedings of the 11th ACM Multimedia Workshop on Multimedia Information Retrieval* (pp. 23 – 30). Berkeley, CA .

Gauvain, J., Lamel, L., & Adda, G. (2002) The LIMSI broadcast news transcription system. *Speech Communication, 37*(1–2), 89–108.

Hauptmann, A., Chen, M. Y., Christel, M., Huang, C., Lin, W. H., Ng, T., Papernick, N., Velivelli, A., Yang, J., Yan, R., Yang, H., & Wactlar, H. D. (2004). Confounded Expectations: Informedia at TRECVID 2004. In *Proceedings of NIST TREC Video Retrieval Evaluation*, Gaithersburg, MD.

Hauptmann, A. G. (2006). Automatic spoken document retrieval. In K. Brown, (Ed.), *Encyclopedia of language and linguistics* 2nd ed. Amsterdam: Elsevier.

Hauptmann, A. G., & Christel, M. G. (2004). Successful approaches in the TREC video retrieval evaluations. In *Proceedings of the 12th annual ACM international conference on Multimedia* (pp. 668–675). New York, NY, USA .

Hauptmann, A. G., Baron, R., Chen, M. Y., Christel, M., Duygulu, P., Huang, C., Jin, R., Lin, W. H., Ng, T., Moraveji, N., Papernick, N., Snoek, C., Tzanetakis, G., Yang, J., Yan, R., & Wactlar, H. (2003a), Informedia at TRECVID 2003: Analyzing and searching broadcast news video. In *Proceedings of NIST TREC Video Retrieval Evaluation*, Gaithersburg, MD.

Hauptmann, A. G., Jin, R., & Ng, T. D. (2003b), Video retrieval using speech and image information. In *Storage and Retrieval for Multimedia Databases 2003, Electronic Imaging '03*, Santa Clara, CA (pp. 148–159).

Hauptmann, A. G., Christel, M., Concescu, R., Gao, J., Jin, Q., Lin, W. H., Pan, J. Y., Stevens, S. M., Yan, R., Yang, J., & Zhang, Y. (2005). CMU Informedia's TRECVID 2005 Skirmishes. In *Proceedings of NIST TREC Video Retrieval Evaluation*, Gaithersburg, MD.

He, J., Li, M., Zhang, H. J., Tong, H., & Zhang, C. (2004), Manifold-ranking based image retrieval. In *Proceedings of the 12th annual ACM international conference on Multimedia* (pp. 9–16). New York, NY, USA

He, X., Ma, W. Y., King, O., Li, M., & Zhang, H. (2002). Learning and inferring a semantic space from user's relevance feedback for image retrieval. In *Proceedings of the tenth ACM international conference on Multimedia* (pp. 343–346). Juan-les-Pins, France.

Hua, X. S., Chen, X. R., Wenyin, L., & Zhang, H. J. (2001). Automatic location of text in video frames. In *Proceedings of the 2001 ACM workshops on Multimedia, Ottawa* (pp. 24–27). Ontario, Canada

Huang, J., Kumar, S., Mitra, M., Zhu, W., & Zabih, R. (1997). Image indexing using color correlograms. In *Proceedings of IEEE Conf. on Computer Vision and Pattern Recognition* (pp. 762–768).

Huang, X., Alleva, F., Hon, H. W., Hwang, M. Y., & Rosenfeld, R. (1993). The SPHINX-II speech recognition system: An overview. *Computer Speech and Language, 7*(2), 137–148.

Huurnink, B. (2005). *AutoSeek: Towards a fully automated video search system*. Master's thesis. Netherlands: University of Amsterdam.

Iyengar, G., Duygulu, P., Feng, S., Ircing, P., Khudanpur, S. P., Klakow, D., Krause, M. R., Manmatha, R., Nock, H. J., Petkova, D., Pytlik, B., & Virga, P. (2005). Joint visual-text modeling for automatic retrieval of multimedia documents. In *Proceedings of ACM Intl. Conf. on Multimedia*.

Jeon, J., Lavrenko, V., & Manmatha, R. (2003). Automatic image annotation and retrieval using cross-media relevance models. In *Proceedings of the 26th annual ACM SIGIR conference on informaion retrieval* (pp. 119–126). Toronto, Canada.

Jin, R., & Hauptmann, A. G. (2002). Using a probabilistic source model for comparing images. In *Proceedings of IEEE Intl. Conf. on Image Processing (ICIP)*, Rochester, NY.

Jing, Y., & Croft, W. B. (1994). An association thesaurus for information retrieval. In *Proceedings of RIAO-94, 4th International Conference "Recherche d'Information Assistee par Ordinateur"* (pp. 146–160). New York, US.

Kennedy, L., Natsev, P., & Chang, S. F. (2005). Automatic discovery of query class dependent models for multimodal search. In *Proceedings of ACM Intl. Conf. on Multimedia* (pp. 882–891). Singapore.

Kraaij, W. (2004). *Variations on Language Modeling for Information Retrieval*. PhD thesis. Netherlands: University of Twente.

Lafferty, J., & Zhai, C. (2003). Probabilistic relevance models based on document and query generation. In *Language Modeling for Information Retrieval, Kluwer International Series on Information Retrieval, vol 13*. Springer.

Lee, T. S. (1996). Image representation using 2D Gabor Wavelets. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 18*(10), 959–971.

Lei, Z., Tasdizen, T., & Cooper, D. (1997). Object signature curve and invariant shape patches for geometric indexing into pictorial databases. In *Proceedings of Multimedia Storage and Archiving Systems II*(pp. 232–243). Dallas, TX.

Leroy, A. M., & Rousseeuw, P. J. (1987). *Robust regression and outlier detection Wiley Series in Probability and Mathematical Statistics*. New York: Wiley.

Lew, M. S., Sebe, N., & Eakins, J. P. (Eds.) (2002). In *Proceedings of Intl. Conf. on Image and Video Retrieval*. London, UK.

Lew, M. S., Sebe, N., Djeraba, C., & Jain, R. (2006). Content-based multimedia information retrieval: State of the art and challenges. *ACM Transactions Multimedia Computing, Communications and Applications, 2*(1), 1–19.

Li, B., & Ma, S. (1994). On the relation between region and contour representation. In *Proc. IEEE Intl. Conf. on Pattern Recognition* (pp. 352–355).

Li, J., Wang, J. Z., & Wiederhold, G. (2000). IRM: integrated region matching for image retrieval. In *Proceedings of the eighth ACM international conference on Multimedia*(pp. 147–156). Marina del Rey, California, United States.

Lienhart, R. (2003). Video OCR: A survey and practitioner's guide. In *Video Mining*. Kluwer Academic Publisher.

Lin, C., Tseng, B., & Smith, J. (2003). VideoAnnEx: IBM MPEG-7 annotation tool for multimedia indexing and concept learning. In *IEEE International Conference on Multimedia and Expo*, Baltimore, MD.

Lu, H., Ooi, B., & Tan, K. (1994). Efficient image retrieval by color contents. In *Proceedings of the 1994 Intl. Conf. on Applications of Databases* (pp. 95–108). Vadstena, Sweden.

Ma, W. Y., & Manjunath, B. S. (1995). A comparison of wavelet transform features for texture image annotation. In *Proceedings of the International Conference on Image Processing (Vol. 2), IEEE Computer Society* (p. 2256). Washington, DC, USA.

MacWorld (2006). Apple preps movie download service. http://www.macworld.co.uk/news/index.cfm?NewsID=13958.

Manjunath, B. S., & Ma, W. Y. (1996). Texture features for browsing and retrieval of image data. *IEEE Transactions on Pattern Analysis and Machchine Intelligence, 18*(8), 837–842.

Marr, D., & Hildreth, E. (1979). Theory of edge detection. In *Proceedings of Royal Society of London Bulletin* (pp. 301–328).

McDonald, K., & Smeaton, A. (2005) A comparison of score, rank and probability-based fusion methods for video shot retrieval. In *International Conference on Image and Video Retrieval(CIVR)*, Dublin, Ireland (pp. 61–70).

Mehtre, B. M., Kankanhalli, M. S., & Lee, W. F. (1997). Shape measures for content based image retrieval: a comparison. *Inf Process Manage, 33*(3), 319–337.

Mitiche, A., & Aggarwal, J. K. (1985). Image segmentation by conventional and information-integrating techniques: a synopsis. *Image and Vision Computing, 3*(2), 50–62.

Nagasaka, A., & Tanaka, Y. (1992). Automatic video indexing and full-video search for object appearances. In *Proceedings of the IFIP TC2/WG 2.6 Second Working Conference on Visual Database Systems II* (pp. 113–127). North-Holland.

Naphade, M. R., & Smith, J. R. (2004). On the detection of semantic concepts at trecvid. In *Proceedings of the 12th annual ACM international conference on Multimedia*(pp. 660–667). New York, NY, USA .

Naphade, M. R., Kristjansson, T., Frey, B., & Huang, T. (1998). Probabilistic multimedia objects (multijects): A novel approach to video indexing and retrieval in multimedia systems. In *Proceedings of IEEE International Conference on Image Processing (ICIP)* (pp. 536–540).

Natsev, A., & Smith, J. R. (2003). Active selection for multi-example querying by content. In *IEEE International Conference on Multimedia and Expo (ICME)*, Baltimore, MA.

Neo, S. Y., Zhao, J., Kan, M. Y., & Chua, T. S. (2006). Video retrieval using high level features: Exploiting query matching and confidence-based weighting. In *Proceedings of the Conference on Image and Video Retrieval (CIVR)* (pp. 370–379). Singapore

Nevatia, R. (1986) Image segmentation. In T. Y. Young, & K.S. Fu (Eds.), *Handbook of pattern recognition and image processing*. San Diego, CA: Academic Press.

Ngo, C. W., Pong, T. C., & Zhang, H. J. (2001). On clustering and retrieval of video shots. In *Proceedings of the ninth ACM international conference on Multimedia* (pp. 51–60). Ottawa, Canada.

Ohanian, P. P., & Dubes, R. C. (1992). Performance evaluation for four classes of texture features. *Pattern Recognition, 25*(2), 819-833.

Pal, N. R., & Pal, S. K. (1993). A review on image segmentation techniques. *Pattern Recognition, 26*, 1277–1294.

Pass, G., & Zabih, R. (1999). Comparing images using joint histograms. *Multimedia System, 7*(3), 234–240.

Perona, P., & Freeman, W. (1998). A factorization approach to grouping. *Lecture Notes in Computer Science, 1406*, 655.

Platt, J. (1999). Probabilistic outputs for support vector machines and comparison to regularized likelihood methods. In A. Smola, P. Bartlett, B. Scholkopf, & D. Schuurmans (Eds.), Advances in Large Margin Classiers. MIT Press.

Ponte, J. M., & Croft, W. B. (1998). A language modeling approach to information retrieval. In *Proceedings of the 21st ACM SIGIR conference on Research and development in information retrieval* (pp. 275–281). Melbourne, Australia.

Puzicha, J., Hofmann, T., & Buhmann, J. (1997). Non-parametric similarity measures for unsupervised texture segmentation and image retrieval. In *Proceedings of IEEE Conf. on Computer Vision and Pattern Recognition* (pp. 267–272).

Qiu, Y., & Frei, H. P. (1993). Concept based query expansion. In *Proceedings of the 16th annual international ACM SIGIR conference*(pp. 160–169). Pittsburgh, Pennsylvania, United States.

Rautiainen, M., Hosio, M., Hanski, I., Varanka, M., Kortelainen, J., Ojala, T., & Seppanen, T. (2004a). TRECVID 2004 experiments at MediaTeam Oulu. In *Proceedings of NIST TREC Video Retrieval Evaluation*, Gaithersburg, MD.

Rautiainen, M., Ojala, T., & Seppanen, T. (2004b). Cluster-temporal browsing of large news video databases. In *IEEE International Conference on Multimedia and Expo (ICME)*, Taipei, Taiwan.

Rickman, R., & Stonham, J. (1996). Content-based image retrieval using color tuple histograms. In *Storage and Retrieval for Image and Video Databases (SPIE)* (pp. 2–7).

Robertson, S. E. (1977). The probability ranking principle in IR. *Journal of Documentation, 33*(4), 294-304.

Robertson, S. E., & Sparck Jones, K. (1977) Relevance weighting of search terms. *Journal of the American Society for Informaiton Science, 27*, 129–146.

Robertson, S. E., & Walker, S. (1994). Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *Proceedings of the 17th ACM SIGIR* (pp. 232–241). Dublin, Ireland .

Robertson, S. E., Walker, S., Hancock-Beaulieu, M., Gull, A., & Lau, M. (1992). Okapi at TREC4. In *Text REtrieval Conference*, Gaithersburg, MD (pp. 21–30).

Rocchio, J. J. (1971). Relevance feedback in information retrieval. In *The SMART Retrieval System: Experiments in Automatic Document Processing*, Prentice Hall, Englewood Cliffs, NJ (pp. 313–323).

Rowe, L. A., & Jain, R. (2004). ACM SIGMM retreat report on future directions in multimedia research. In *Proceedings of ACM Multimedia.*

Rui, Y., She, A., & Huang, T. (1996). Modified Fourier descriptors for shape representation – A practical approach. In *Proceedings of First International Workshop on Image Databases and Multimedia Search*, Amsterdam, The Netherlands.

Rui, Y., Huang, T., & Mehrotra, S. (1997a). Content-based image retrieval with relevance feedback in MARS. In *Proc. IEEE Intl. Conf. on Image Processing* (pp. 815–818).

Rui, Y., Huang, T. S., & Chang, S. F. (1997b). Image retrieval: Past, present, and future. In *International Symposium on Multimedia Information Processing.*

Salton, G. (1989). *Automatic text processing*. Addison-Wesley.

Sarkar, S., & Boyer, K. L. (1996). Quantitative measures of change based on feature organization: Eigenvalues and eigenvectors. In *IEEE Computer Vision and Pattern Recognition(CVPR)* (pp. 478–483).

Sato, T., Kanade, T., Hughes, E., & Smith, M. (1998). Video OCR for digital news archives. In *IEEE Workshop on Content-Based Access of Image and Video Databases(CAIVD'98)* (pp. 52 – 60).

Schmid, C., & Mohr, R. (1997) Local grayvalue invariants for image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 19*(5), 530–535.

Shaw, J. A., & Fox, E. A. (1994). Combination of multiple searches. In *Text REtrieval Conference*, Gaithersburg, MD.

Shi, J., & Malik, J. (1998). Motion segmentation and tracking using normalized cuts. In *Proceedings of Intl. Conf. on Computer Vision (ICCV)* (pp. 1154–1160).

Shi, J., & Malik, J. (2000). Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 22*(8), 888–905.

Singhal, A., Buckley, C., & Mitra, M. (1996). Pivoted document length normalization. In *Proceedings of the 19th ACM SIGIR, Zurich* (pp. 21–29). Switzerland

Sivic, J., & Zisserman, A. (2003). Video google: A text retrieval approach to object matching in videos. In *Proceedings of the Ninth IEEE International Conference on Computer Vision*, p. 1470.

Smeaton, A., & Over, P. (2003). TRECVID: Benchmarking the effectiveness of information retrieval tasks on digital video. In *Proceedings of the Intl. Conf. on Image and Video Retrieval* (pp. 19–27).

Smeaton, A., Over, P., & Kraaij, W. (2006). Evaluation campaigns and TRECVid. In *Proceedings of the 8th ACM international workshop on Multimedia information retrieval* (pp. 321–330).

Smeulders, A. W. M., Worring, M., Santini, S., Gupta, A., & Jain, R. (2000). Content-based image retrieval: the end of the early years. *IEEE Transactions on Pattern Analysis Machine Intelligence, 12*, 1349–1380.

Smith, J. R., & Chang, S. F. (1996a). Automated binary texture feature sets for image retrieval. In *Proceedings of the IEEE International Conference on Acoustics*, Speech, and Signal Processing (pp. 2239–2242).

Smith, J. R., & Chang, S. F. (1996b). Tools and techniques for color image retrieval. In *Storage and Retrieval for Image and Video Databases (SPIE)* (pp. 426–437).

Smith, J. R., & Chang, S. F. (1996c). Visualseek: A fully automated content-based image query system. In *ACM Multimedia* (pp. 87–98).

Smith, J. R., Lin, C. Y., Naphade, M. R., Natsev, P., & Tseng, B. (2002). Advanced methods for multimedia signal processing. In *Proceedings of Intl. Workshop for Digital Communications*, Capri, Italy.

Snoek, C., & Worring, M. (2005). Multimodal video indexing: A review of the state-of-the-art. *Multimedia Tools Application, 25*(1), 5–35.

Snoek, C., Worring, M., Geusebroek, J., Koelma, D., & Seinstra, F. (2004), The MediaMill TRECVID 2004 semantic viedo search engine. In *Proceedings of NIST TREC Video Retrieval Evaluation*, Gaithersburg, MD.

Snoek, C., Worring, M., & Smeulders, A. (2005). Early versus late fusion in semantic video analysis. In *Proceedings of ACM Intl. Conf. on Multimedia*(pp. 399–402). Singapore.

Srikanth, M., Bowden, M., & Moldovan, D. (2005). LCC at trecvid 2005. In *Proceedings of NIST TREC Video Retrieval Evaluation*, Gaithersburg, MD.

Stricker, M. A. (1994). Bounds for the discrimination power of color indexing techniques. In *Storage and Retrieval for Image and Video Databases (SPIE)* (pp. 15–24).

Stricker, M. A., & Orengo, M. (1995). Similarity of color images. In *Storage and Retrieval for Image and Video Databases (SPIE)* (pp. 381–392).

Su, Z., Li, S., & Zhang, H. (2001). Extraction of feature subspaces for content-based retrieval using relevance feedback. In *Proceedings of the ninth ACM international conference on Multimedia* (pp. 98–106). Ottawa, Canada.

Swain, M. J., & Ballard, D. H. (1991). Color indexing. *International Journal on Computer Vision, 7*(1), 11–32.

Szummer, M., & Picard, R. (2002). Indoor-outdoor image classification. In *IEEE International Workshop in Content-Based Access to Image and Video Databases*, Bombay, India.

Thyagarajan, K., Nguyen, J., & Persons, C. (1996). A maximum likelihood approach to texture classification using wavelet transform. In *Proceedings of the International Conference on Image Processing (ICIP)* (pp. 640–644).

Turtle, H. R. (1991). *Inference Networks for Document Retrieval*. PhD thesis, University of Massachusetts.

Tuytelaars, T., & van Gool, L. J. (1999). Content-based image retrieval based on local affinely invariant regions. In *Proceedings of the Third International Conference on Visual Information and Information Systems*(pp. 493–500). Springer-Verlag, London, UK.

Vogt, C. C., & Cottrell, G. W. (1999). Fusion via a linear combination of scores. *Information Retrieval, 1*(3), 151–173.

Volkmer, T., & Natsev, A. (2006). Exploring automatic query refinement for text-based video retrieval. In *IEEE International Conference on Multimedia and Expo (ICME)*, Toronto, ON (pp. 765–768).

Wactlar, H., Christel, M., Gong, Y., & Hauptmann, A. G. (1999) Lessons learned from the creation and deployment of a terabyte digital video library. *IEEE Computer, 32*(2), 66–73.

Westerveld, T. (2004). *Using generative probabilistic models for multimedia retrieval*. PhD thesis, CWI, Centre for Mathematics and Computer Science.

Westerveld, T., & de Vries. A. (2004). Multimedia retrieval using multiple examples. In *International Conference on Image and Video Retrieval (CIVR)*, Dublin, Ireland (pp. 344–352).

Westerveld, T., Ianeva, T., Boldareva, L., de Vries, A. P., & Hiemstra, D. (2003). Combining infomation sources for video retrieval: The lowlands team at TRECVID 2003. In *Proceedings of NIST TREC Video Retrieval Evaluation*, Gaithersburg, MD.

White, R. W., Jose, J. M., & Ruthven, I. (2006). An implicit feedback approach for interactive information retrieval. *Information Processing and Management, 42*(1), 166–190.

Wu, Y., Chang, E. Y., Chang, K. C. C., & Smith, J. R. (2004). Optimal multimodal fusion for multimedia data analysis. In *Proceedings of the 12th annual ACM international conference on Multimedia* (pp. 572–579). New York, NY, USA.

Xu, J., & Croft, W. B. (2000). Improving the effectiveness of information retrieval with local context analysis. *ACM Transactions on Information System, 18*(1), 79–112.

Yan, R. (2006). *Probabilistic models for combining diverse knowledge sources in multimedia retrieval*. PhD thesis. Pittsburgh, PA: School of Computer Science, Carnegie Mellon University.

Yan, R., &Hauptmann, A. G. (2006). Probabilistic latent query analysis for combining multiple retrieval sources. In *Proceedings of the 29th annual international ACM SIGIR conference on information retrieval* (pp. 324–331). Seattle, Washington, USA.

Yan, R., Yang, J., & Hauptmann, A. G. (2004). Learning query-class dependent weights in automatic video retrieval. In *Proceedings of the 12th annual ACM international conference on Multimedia* (pp. 548–555). New York, NY, USA.

Yang, H., Chaisorn, L., Zhao, Y., Neo, S. Y., & Chua, T. S. (2003). VideoQA: Question answering on news video. In *Proceedings of the 11th ACM Multimedia* (pp. 632–641). Berkeley, CA, USA.

Yang, J., Chen, M. Y., & Hauptmann, A. G. (2004) Finding person X: Correlating names with visual appearances. In *Proceedings of the Intl. Conf. on Image and Video Retrieval (CIVR)*(pp. 270–278). Dublin, Ireland.

Yuan, J., Xiao, L., Wang, D., Ding, D., Zuo, Y., Tong, Z., Liu, X., Xu, S., Zheng, W., Li, X., Si, Z., Li, J., Lin, F., &Zhang, B. (2005). Tsinghua university at TRECVID 2005. In *Proceedings of NIST TREC Video Retrieval Evaluation*, Gaithersburg, MD.

Zahn, C., & Roskies, R. (1972). Fourier descriptors for plane closed curve. *IEEE Transactions on Computers, 21*, 269–281.

Zhai, C., & Lafferty, J. (2001). A study of smoothing methods for language models applied to ad hoc information retrieval. In *Proceedings of the 24th ACM SIGIR conference on Research and development in information retrieval*(pp. 334–342). New Orleans, Louisiana, United States.

Zhai, Y., Liu, J., Cao, X., Basharat, A., Hakeem, A., Ali, S., Shah, M., Grana, C., & Cucchiara, R. (2005). Video understanding and content-based retrieval. In *Proceedings of NIST TREC Video Retrieval Evaluation 2005*, Gaithersburg, MD.

Zhang, H. J., Smoliar, S. W., Wu, J. H., & Low, C. Y. (1994). Development of a video database system. *SIGOIS Bull, 15*(1), 9.