# Video Search Reranking through Random Walk over Document-Level Context Graph

Winston H. Hsu
Graduate Institute of Networking and Multimedia
National Taiwan University
Taipei, Taiwan
winston@csie.ntu.edu.tw

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

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

## ABSTRACT

Multimedia search over distributed sources often result in recurrent images or videos which are manifested beyond the textual modality. To exploit such contextual patterns and keep the simplicity of the keyword-based search, we propose novel reranking methods to leverage the recurrent patterns to improve the initial text search results. The approach, *context reranking*, is formulated as a random walk problem along the *context graph*, where video stories are nodes and the edges between them are weighted by multimodal contextual similarities. The random walk is biased with the preference towards stories with higher initial text search scores – a principled way to consider both initial text search results and their implicit contextual relationships. When evaluated on TRECVID 2005 video benchmark, the proposed approach can improve retrieval on the average up to 32% relative to the baseline text search method in terms of story-level Mean Average Precision. In the people-related queries, which usually have recurrent coverage across news sources, we can have up to 40% relative improvement. Most of all, the proposed method does not require any additional input from users (e.g., example images), or complex search models for special queries (e.g., named person search).

**Categories and Subject Descriptors:** H.3.3 [Information Search and Retrieval]: Retrieval models

**General Terms:** Algorithms, Performance, Experimentation

**Keywords:** Video Search, Multimodal Fusion, Power Method

## 1. INTRODUCTION

Image and video retrieval has been an active research area thanks to the continuing growth of online video data, personal video recordings, 24-hour broadcast news videos, etc. The phenomenal success in WWW search has also helped attract increasing interest in investigating new solutions in video search.


*MM'07,* September 23–28, 2007, Augsburg, Bavaria, Germany.



(a) Arabic web news (BBC) (b) English web news (Associated Press) (c) Chinese web news (Chinatimes)

(d) English broadcast news video (CBS) with the transcripts

**Figure 1: Examples of a broadcast news video (d) and three web news (a-c) of different languages covering the same topic "Pope discusses his remarks on Islam," collected on September 17, 2006. The images of Pope Benedict XVI are widely used (in near-duplciates) over all the news sources of the same topic. Aside from the text transcripts or web text tokens, the visual duplicates provide another similarity link between broadcast news videos or web news and help cross-domain topic threading, or boost video search performance.**

Current video search approaches are mostly restricted to text-based solutions which process keyword queries against text tokens associated with the video, such as speech transcripts, closed captions, and recognized video text (OCR). However, such textual information may not necessarily come with the image or video sets. The use of other modalities such as image content, audio, face detection, and high-level concept detection has been shown to improve upon text-based video search systems [11, 7, 4, 6, 28] . Such multimodal systems improve the search performance by utilizing multiple query example images, specific semantic concept detectors (e.g., search for cars), or highly-tuned retrieval models for specific types of queries (e.g., use face/speaker ID for searching named persons). Fig. 1 shows an example query of "Pope," which may be answered by using multimodal cues extracted from the face photos, speech transcripts, as well as some predefined concepts related to activities or locations.

However, there are difficulties in applying the multimodal search methods mentioned above. It is quite difficult for the users to acquire example images as query inputs. Retrieval

by matching semantic concepts, though promising, strongly depends on availability of robust detectors and usually requires large amounts of data for training the detectors. Additionally, it is still an open issue whether concept models for different classes of queries may be developed and proved effective across multiple domains. Additionally, based on the observations in the current retrieval systems [2], most users expect searching images and videos simply through a few keywords. Therefore, incorporation of multimodal search methods should be as transparent and non-intrusive as possible, in order to keep the simple search mechanism preferred by typical users today.

Based on the above observations and principles, we propose to conduct semantic video search in a *reranking* manner which automatically reranks the initial text search results based on the "recurrent patterns" or "contextual[1] patterns" between videos in the initial search results. Such patterns exist because of the common semantic topics shared between multiple videos, and additionally, the common visual content (i.e., people, scenes, events, etc.) used. For example, Fig. 1 shows a broadcast news video and three web news articles in different languages (e.g., Arabic, English, and Chinese) covering the same topic "Pope discusses his remarks on Islam." Apparently, the visual near-duplicates of Pope Benedict XVI are used over all the news sources reporting the same topic. Such visual duplicates provide strong links between broadcast news videos or web news articles and help cross-domain information exploitation.

Such recurrent images or videos are commonly observed in image search engines (e.g., Yahoo! or Google) and photo sharing sites (e.g., Flickr). In [15], we had analyzed the frequency of such recurrent patterns (in terms of visual duplicates) for cross-language topic tracking – a large percentage of international news videos share re-used video clips or near duplicates.

Video reranking has been explored in some prior works [16, 7, 28, 10]. [16] specifically explored recurrent patterns in the initial search results and used an Information Bottleneck principle to develop a reranking method. It achieved significant performance gains when applied to the retrieval of video shots. However, such a method is restricted because of its dependence on the representation and occurrence frequency estimation at the shot level, rather than the semantic level. Typically, a semantic document of videos contains multiple images – e.g., a broadcast news story has multiple shots (as shown in Fig. 1-(d)) and a web page may include multiple photos or video clips (as shown in Fig. 1-(a), (b), and (c)) in addition to associated textual data in the same document. Therefore, in this paper, we propose to use multimodal similarities between semantic video documents (e.g., news stories) in exploiting the underlying contextual patterns and developing a new reranking method, *context reranking*. Fig. 2 shows another example motivating the story-level reranking approach. An initial text query retrieves video stories with the keywords "Tony Blair." However, there are still certain relevant stories not retrieved (or assigned low text relevance scores) due to the lack of keyword annotations associated with them. The multimodal contextual links (e.g., visual duplicates, text, etc.) may be used to link such missing stories to the initial text queries and further improve the search accuracy.

[1]The meanings of "context" are usually application-dependent [8]. Here, we refer to context as those attributes describing *who*, *where*, *when*, *what*, etc., shared by documents forming the recurrent patterns.

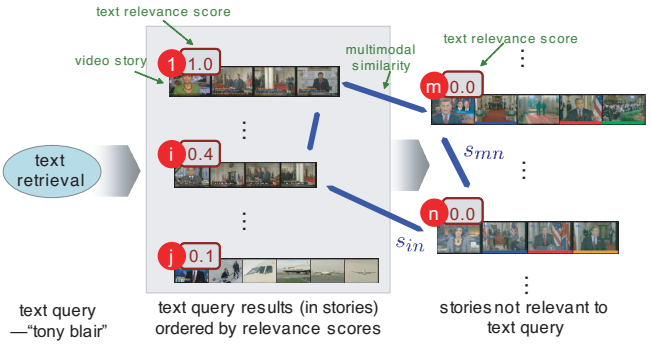


**Figure 2: Example of a video search that benefits from multimodal story-level similarity on a large-scale video database, even with unreliable text ASR/MT transcripts. Though not relevant to the text query, certain stories can be boosted due to their closeness to some relevant text-query stories by the multimodal similarity (shown in the right panel) consisting of text, visual duplicates, etc.**

Due to the lack of explicit links between video stories, context reranking is formulated as a random walk over the *context graph* whose nodes represent documents in the search set. These nodes are connected by the edges weighted with pair-wise multimodal contextual similarities. The stationary probability of the random walk is used to compute the final scores of video stories after reranking. The random walk is biased with the preference towards the stories with higher initial text search scores – a principled way to combine both initial text search results and their implicit contextual relationships. Our contextual reranking method is in principle inspired by the page ranking techniques widely used in Web search [23]. Our innovations lie in the use of the multimodal similarities and the novel integration of the initial text search and reranking results via a rigorous random walk framework. We also investigate the optimal weights for combining the text search scores and the multimodal reranking for different classes of video queries.

Our experiments (cf. section 4) show that the proposed approach can improve the baseline text retrieval up to 32% in story-level Mean Average Precision (MAP[2]). Such results are remarkable since no additional advanced methods are used, like query expansion, specific visual concept detectors, or face/speaker detection. Furthermore, for people-related queries, which usually have recurrent coverage across news sources, our experiments show up to 40% relative improvement in story-level MAP. Through parameter sensitivity tests, we also discovered that the optimal text vs. visual weight ratio for reranking initial text search results is 0.15 to 0.85. It is encouraging and intuitive since, as illustrated in Fig. 2, visual modality plays an important role in defining the contextual similarity between video stories.

The paper is organized as follows. We formulate the reranking problem in section 2. The multimodal contextual similarity is introduced in section 4.3. Based on these methods, the video reranking approach is proposed in section 3.

[2]MAP: mean average precision, a search performance metric used in TRECVID. See more details in Section 4.1.

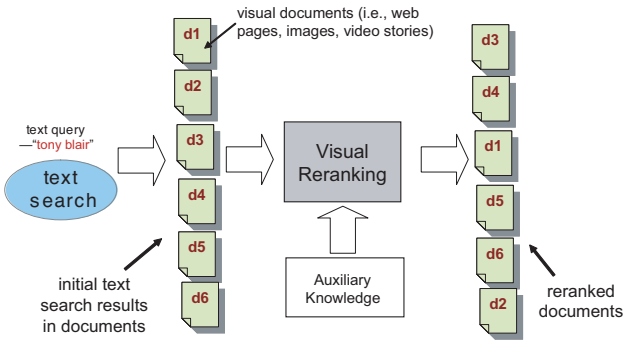


**Figure 3: The illustration of the proposed visual reranking process which tries to improve the visual documents (i.e., web pages, images, videos, etc.) from initial text search results.**

The experiments are conducted in section 4. We present further discussion and review of related work in section 6, and conclusions in section 7.

## 2. VISUAL RERANKING

Assuming we have $n$ visual documents (i.e., web pages, images, video stories) $\{d_1, d_2, ..., d_n\}$, from initial text search results, as illustrated in Fig. 3. The visual reranking process is used to improve the search accuracy by reordering the visual documents based on the multimodal cues extracted from the initial text search results and the auxiliary knowledge, if available. The auxiliary knowledge can be the extracted features from each visual document or the multimodal similarities between them and will be discussed later in the following section.

To this end, pseudo-relevance feedback (PRF) [5, 7, 28, 21], is one such tool which has been shown effective in improving initial text search results in both text and video retrieval. PRF is initially introduced in [5], where a significant fraction of top-ranked documents are assumed to be relevant and are then used to build a model for reranking the search result set. This is in contrast to relevance feedback where users explicitly provide feedback by labeling the results as positive or negative.

The same concept has been implemented in video retrieval. In [7], authors used the textual information in the top-ranking shots to obtain additional keywords to perform retrieval and rerank the initial shot lists. The experiment was shown to improve MAP from 0.120 to 0.124 in the TRECVID 2004 video search task [1]. In [28], authors sampled the pseudo-negative images from the lowest rank of the initial query results, used the query videos and images as the positive examples, and retrieval was formulated as a classification problem. This resulted in improvement in the search performance from MAP 0.105 to 0.112 in TRECVID 2003. The authors of [10] used the Google image search return set as (pseudo-)positives and utilized a parts-based approach to learn the object model and then used that to rerank the initialsearch images. The object model was selected, through a scoring function, among a large number (∼100) of hypothesized parts-based models, which are very time consuming. Furthermore, their experiment was limited to image queries of simple objects such as bottles, cars, etc., instead of natural language queries as those in TRECVID.

Reranking in video search has been pursued even further in more recent works. In [27], the authors provide a robust basis for relevance feedback in interactive search by introducing the notion of video "threads," which are links between video shots based on any number of criteria, including visual appearance, common concept detection scores, and temporal adjacency. As users browse down through a list of video search results, when relevant shots are found, similar shots, along with any given video thread, can also be explored, increasing the likelihood of finding more relevant shots. In [12], the authors incorporate a similar approach, which uses temporal adjacency to feed back into the search result browsing interface. Both of these methods exploit the visual recurrence present in large parallel news corpora and the likelihood of recurrence within a single news story (indirectly via the temporal adjacency); however, these relationships are all exploited in an interactive interface, which requires explicit feedback from the searcher. In this work, we are interested in exploiting such robust connections automatically via a pseudo-relevance feedback-like approach.

In [4] many different kinds of automatic video search queries can be leveraged to rerank and refine each other. Specifically, text search is conducted over speech transcripts using both story segmentation and shot segmentation and the results of each are used to refine and rerank each other. Similarly, the results of a search based on a mapping from text keywords to pre-trained concept detection scores can be used as a basis for reranking other search methods. "Reranking" in this case is more like a fusion between individual search methods and does not specifically employ the visual recurrence in news as a pseudo-relevance feedback mechanism for mining visual patterns in initial search results.

## 3. VIDEO SEARCH RERANKING VIA RANDOM WALK

### 3.1 Random Walk on Multimodal Story-Level Similarities

We formulate the context reranking solution as a random walk over the *context graph*, where stories are nodes and the edges between them are weighted by multimodal contextual similarities, as illustrated in Fig. 4 and defined in Eqn. 5. We use the *context ranking score*, the stationary probability of the random walk over the context graph, to represent the search relevance scores. The purpose is to rerank the initial text search results by their implicit context – the multimodal pair-wise similarities – residing in the video story collections.

Assuming we have $n$ nodes in the random walk process. A node might correspond to a Web page in the text retrieval problem or a story in the broadcast news video. We will discuss and compare various strategies of determining the number of nodes (video stories) in section 3.3.

A stochastic (transition) matrix $\mathbf{P} \equiv [p_{ij}]_{n\times n} = [p(j|i)]_{n\times n}$ is used to govern the transition of a random walk process. $p_{ij}$ is the probability that transition from state $i$ to state $j$ occurs. The state probability at time instance $k$ is $\mathbf{x}_{(k)} \equiv \left[p_{(k)}(i)\right]_{n\times 1}$, a column vector of the probabilities residing in the $n$ nodes at the instance. The stationary probability $\mathbf{x}_\pi \equiv \lim_{k\to\infty} \mathbf{x}_{(k)}$ is the state probability of the random walk process as the time instance proceeds to infinity if the convergence conditions are satisfied.

In this context, we consider both the multimodal simi-

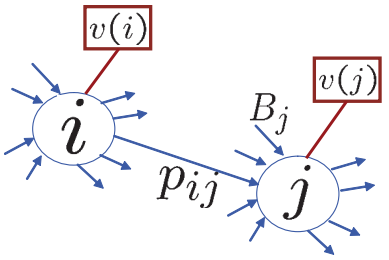


**Figure 4: Example of a context graph for random walk over nodes (stories); $i$ and $j$ are node index with their original text search scores $v(i)$ and $v(j)$; $p_{ij}$ is the transition probability from node $i$ to $j$; $B_j$ are edges connected to node $j$.**

larities or transition probabilities $\mathbf{P}$ between stories and the original (normalized) text search scores $\mathbf{v}$, or called *personalization vector* [19]. In this framework, the state probability $x_{(k)}(j)$ of node $j$ at time instance $k$ is:

$$x_{(k)}(j) = \alpha \sum_{i \in B_j} x_{(k-1)}(i) p_{ij} + (1-\alpha) v(j), \qquad (1)$$

where $B_j$ is the set of edges connecting back to node $j$, $p_{ij}$ is the contextual transition probability from story $i$ to $j$, and $\alpha \in [0, 1]$ linearly weights two terms.

Eqn. 1 is actually an interesting interpretation of the random walk process exemplified in Fig. 4 and 2. Intuitively, $x_{(k)}(j)$ is parameterized by its neighboring nodes $B_j$ at time instance $k-1$ and its own initial text scores $v(j)$. Both are then linearly fused with weights $\alpha$ and $1-\alpha$ respectively. For the first term of Eqn. 1, we consider not only the state probabilities of its neighbors $B_j$ but also their corresponding transition probabilities – how possible it is to reach node $j$. The second term is the initial text score for node $j$. Such linear fusion considers the state probabilities (or search context relevances) of its neighbors and its initial text scores. The linear fusion is commonly used in multimodal search and concept detection research [9] and has been shown to be effective.

The relationship in Eqn. 1 is updated recursively until all nodes in the graph converge. For each node, the new search relevance score is its stationary probability, if it exists. For example, according to Eqn. 1, the stationary probability of node $j$ is

$$x_\pi(j) = \alpha \sum_{i \in B_j} x_\pi(i) p_{ij} + (1-\alpha) v_j. \qquad (2)$$

Naturally, we have $\sum_j x_\pi(j) = 1$. If we set $\mathbf{E} = \mathbf{e}\mathbf{v}^T$ (where $\mathbf{e}$ is an $n$-dimensional column vector with all 1) and apply some algebraic manipulations, then we can get

$$\begin{aligned} \mathbf{x}_\pi^T \equiv [x_\pi(j)]_{1\times n} &= \alpha \mathbf{x}_\pi^T \mathbf{P} + (1-\alpha)\mathbf{x}_\pi^T \mathbf{E} \\ &= \mathbf{x}_\pi^T [\alpha \mathbf{P} + (1-\alpha)\mathbf{E}] = \mathbf{x}_\pi^T [\mathbf{P}']. \end{aligned} \qquad (3)$$

The solution of $\mathbf{x}_\pi$ is apparently one of the eigenvectors of $[\mathbf{P}']^T$. As shown in [19], it is exactly the (normalized) dominant eigenvector, the one corresponding to the largest absolute eigenvalue of $[\mathbf{P}']^T$. Meanwhile, Eqn. 3 has almost the same formulation as the PageRank algorithm, where the stationary probability can be derived through an efficient algorithm, the Power Method, [23, 19], which is briefly introduced in the following section.

Interestingly, PageRank considers the links between web pages to quantitatively measure the "quality" of the web pages instead of considering the text search "relevance" only. In our approach, rather than constructing those "contextual links" manually, we adopt the soft multimodal similarities between the story pairs, which are computed automatically.

We need to derive the context graph transition probability from the raw story-level affinity matrix $\mathbf{S} = [s_{ij}]_{n\times n}$, as defined in Eqn. 5 (cf. section 4.3). One intuitive way is to apply the normalization process and ensure that each row sums to 1. That is,

$$p_{ij} = \frac{s_{ij}}{\sum_k s_{ik}}.$$

A random walk process will more likely jump to the nodes with higher (contextual) similarities. Similar to the PageRank algorithm, we need to handle the same dangling problem and just set those all zero rows of affinity matrix $\mathbf{S}$ as $\mathbf{e}^T$. The affinity matrix $\mathbf{S}$ is composed of pair-wise similarities between stories. The story-level similarity consists of visual duplicates and text as depicted in Eqn. 5. In our following experiments, variant modality weights between visual duplicates and text are tested experimentally and will be further discussed in section 5.4.1.

## 3.2 Power Method

To calculate the dominant eigenvector of a stochastic matrix $\mathbf{A}$ (i.e. $[\alpha\mathbf{P} + (1-\alpha)\mathbf{E}]^T$ in Eqn. 3), we adopted the Power Method [18] which iteratively applies $\mathbf{x}_{(k)} = \mathbf{A}\mathbf{x}_{(k-1)}$ until $\mathbf{x}_{(k)}$ converges. Note that $\mathbf{x}_{(k)}$ needs to be normalized by its 1-norm ($||\cdot||_1$) to make sure that it adds up to 1. Applying this method iteratively can result in convergence to the stationary probability $\mathbf{x}_\pi$.

Matrix $\mathbf{E}$ in Eqn. 3 is important in avoiding the case that the random walk process is trapped in local optimal solutions due to the sparseness of transition matrix P. It further helps ensure the random walk process will converge [19]. Our experiments (cf, section 5.3) have shown very good convergence property of such iterative processes.

## 3.3 Improving Text Search with Context Ranking

The combination of the context graph formulation (section 3.1) and the Power Method for random walk (section 3.2) provides a powerful and flexible framework for reranking initial results from video search, and fusion of search results by different modalities. There are many interesting dimensions for consideration when we apply the framework in practice: (1) Which video stories form the nodes of the context graph? The whole set of video stories or those with textually relevant scores only? (2) How should we use the initial text search results? Fused with the context ranking scores or acting as the prior in the personalization vector $\mathbf{v}$ for the random walk process? To maximize the performance of these promising methods and modalities, we test several combinations of the approaches in this section. The experiment results are shown in section 5.1.

- Full-Ranking (**FR**): In *full-ranking*, all video stories in the database are used to construct the context graph for the random walk framework. The personalization

vector is $\mathbf{v} = \mathbf{e}/n$, uniform over all stories. The stationary probability, then, is used as the context ranking score. The assumption behind FR is that those stories having the most recurrences will be given higher context ranking scores. This is not the best setup since it does not utilize the initial text search scores but it does provide a baseline for comparison.

- Full-Ranking and Text Score (**FRTS**): The context rank scores, derived in FR, are then fused with the the initial text scores by averaging the two scores. We use simple averaging here rather than optimize varying weights, which generated comparable performance in prior studies [9].

- Full-Ranking with Text Prior (**FRTP**): Same as FR, except that the initial text score is used as the personalization vector $\mathbf{v}$ or the random walk prior. The second term, $v(j)$, in Eqn. 1 is no longer equal across all stories but biased by the initial text scores. Such a prior will place more preference towards those stories with higher initial text search scores. The effect is modulated by the linear weighting factor $\alpha$.

- Partial Ranking (**PR**): In *partial ranking*, only stories in the database that have initial text search scores higher than some threshold (i.e., 0) are used to construct the context graph. The multimodal similarity graph $\mathbf{S}$ is first built based on this subset of stories relevant to the query terms. A uniform personalization vector $\mathbf{v}$ over the subset of stories, as in FR, is used. The assumption is to consider those text-query-relevant stories only and then order them by the recurrence frequency – more recurrent stories gain more relevance, or higher stationary probability, from the partial context graph.

- Partial-Ranking and Text Score (**PRTS**): Same as PR, except the context ranking score is further fused with the initial text scores by averaging the two.

- Partial-Ranking with Text Prior (**PRTP**): Same as PR, except that the initial text search score is used as the personalization vector $\mathbf{v}$ for the random walk prior. The assumption is to consider the recurrent patterns within the text-query-relevant stories only and biased with the prior of the initial text search scores.

- Partial-Ranking with Text Prior and Link Reduction by $K$ Nearest Neighbors (**PRTP-KNN**): In the previous method PRTP, the context graph built over the subset of relevant stories is a complete graph or composed of all edges between the nodes in the graph. The authors of [24] had discovered that selecting the top-$K$ most significant edges between nodes can further boost the random walk performance. Likewise, in this method, we select the top-$K$ edges, originating from each node and having highest similarities. Note that to discover the effects of $K$, we experiment with varying values of $K$ in the following experiments.

In the 2nd column, we added labels (TS and TP) to indicate different ways of incorporating text search scores with the multimodal context graph. For TS, the text scores and the reranking scores are simply averaged; for TP, the text scores are used as the random walk priors.

# 4. EXPERIMENTAL SETUPS

## 4.1 Data set

The experiments are conducted over the TRECVID 2005 data set [1], which contains 277 international broadcast news video programs and accumulates 170 hours of videos from 6 channels in 3 languages (Arabic, English, and Chinese). The time span is from October 30 to December 1, 2004. The ASR and machine translation (MT) transcripts are provided by NIST [1].

In [16], the authors have shown that an effective approach to text searches over ASR/MT transcripts from international broadcast news videos is to use the transcripts from the entire story. This makes sense since the true semantic relationships between images and the transcripts exist at the story level: if a concept is mentioned in the text it is likely to appear in the images somewhere within the same story, but unlikely to appear in the next story or the previous one. Story boundaries can be extracted with reasonable accuracy automatically by analyzing the visual characteristics of the image content and the speech of the anchorperson [14].

The search ground truth provided in TRECVID 2005 is at the shot level. Since our primary interest is at the story level, we have to convert the shot-level ground truth into story-level ground truth. A story is considered positive if it contains at least one shot labeled as positive for a particular query. It is an "OR" operation among shots in the same story. This is intuitive since browsing stories at the story level is more natural and a story is a semantically coherent unit; a story should be labeled as positive once a target object or shot is discovered in the story. For example, a user might be interested in news related to Chinese President Hu Jintao and thus any story that visually shows the presence of the person in any shot should be considered relevant.

For measuring performance, we adopted non-interpolated average precision (AP), which approximates the area under a (non-interpolated) recall/precision curve. Since AP only shows the performance of a single query, we use mean average precision (MAP), simply the mean of APs for multiple queries, to measure average performance over sets of different queries in a test set. See more explanations in [1].

## 4.2 Story-level search baselines

Natural language processing tools such as stemming, part-of-speech tagging, etc, are first applied to the ASR/MT transcripts provided by NIST. We conduct the experiments on two text-based video search sets, "text-okapi" and "text-svm." The former is from the baseline text search approach based on Okapi method [16]. The latter is motivated by multi-bag support vector machines (MB-SVM)[3] in [21]. We sample the ASR/MT transcripts from stories associated with the example video clips as positive examples and randomly sample other stories in the database for pseudo-negatives. A discriminative model (e.g., SVM) is trained to classify other stories in the test set. The process is repeated several times (with different random samples of negative data) and the (positive) distances to the margin plane of these SVM models are then fused to compute the final relevance scores. In TRECVID 2005 data set, the "text-svm" approach signifi-

[3]Similar to the supervised training in [21], this method requires users to provide example text transcripts to build text-based search models. Here, we still include it as an alternative baseline for performance comparison.

| methods | text fusion | text-okapi MAP | text-okapi % | text-svm MAP | text-svm % |
|---|---|---|---|---|---|
| initial text search | - | 0.204 | - | 0.274 | - |
| Full Ranking (FR) | - | 0.109 | -46.8 | 0.109 | -60.3 |
| Full Ranking and Text Score (FRTS) | TS | 0.238 | 16.5 | 0.302 | 10.2 |
| Full Ranking with Text Prior (FRTP) | TP | 0.210 | 2.8 | 0.280 | 2.1 |
| Partial Ranking (PR) | - | 0.196 | -4.1 | 0.240 | -12.3 |
| Partial Ranking and Text Score (PRTS) | TS | 0.255 | 24.5 | 0.309 | 12.9 |
| **Partial Ranking with Text Prior (PRTP)** | TP | 0.271 | **32.5** | 0.333 | **21.6** |
| PRTP-KNN (with best $K$) | TP | 0.271 | 32.5 | 0.333 | 21.6 |

**Table 1: The performance (MAP at depth 20 and $\alpha = 0.8$ in Eqn. 3) and relative improvements (%) from the initial text search results at different methods (cf. section 3.3). Note that in PRTP-KNN the best results among variant $K$ (number of neighbors) are shown. In the second column, we added labels to indicate different ways of incorporating text search scores with the multimodal context graph. For *TS*, the text scores and the reranking scores are simply averaged; for *TP*, the text scores are used as the random walk priors.**

cantly outperforms the "text-okapi" (cf. Tab. 1). See more explanations in [6].

## 4.3 Story-level feature representations

To construct multimodal similarities between story pairs and further support video search ranking, two fundamental issues need to be addressed – (1) representation of each story and (2) measurement of similarity between story pairs. In this section, we first describe the text processing techniques to extract cue word clusters and then present parts-based near-duplicate detection for visual similarity.

### 4.3.1 Text – cue word clusters

We represent the text modality of each story by compact "cue word clusters" or "pseudo-words" [26]. The text transcripts for each story are from the automatic speech recognition (ASR) and machine translation (MT) transcripts included in the TRECVID 2005 data set [1][4]. The first text processing step involves stemming the ASR and MT tokens and removing the stop words, resulting in a total of 11562 unique words. Then a mutual information (MI) approach is used to select the top 4000 informative words (suggested in [26]) based on the MI between the stories and words. Specifically, given a set of stories $D$ over the word set $O$, the MI between stories and words can be computed as $I(O;D) = \sum_{o \in O} I(o)$, where $I(o)$ represents the contribution of word $o$ to the total MI $I(O;D)$.

$$I(o) \equiv p(o) \sum_{d \in D} p(d|o) \log \frac{p(d|o)}{p(d)}, \quad (4)$$

where the probability terms needed above can be easily computed from the co-occurrence table between words and stories.

These words are further grouped into 120[5] cue word clusters (or pseudo-words) using the Information Bottleneck (IB) principle [26]. Words in the same cue word cluster are associated with the same semantic – for example, {*insurgent*, *insurgents*, *iraq*, *iraqis*, *iraqi*, *marine*, *marines*, *troops*} or {*budget*, *capitol*, *politics*, *lawmakers*, *legislation*, *legislative*, *reform*}. Later each story is represented as a 120-dimensional pseudo-word frequency vector.

[4]The experiments are conducted over the 2580 automatically-decteced stories in the "test" set.

[5]The number is selected based on the empirical experiment in our prior work [15].

The story-level similarity $\psi_t(d_i, d_j)$ is computed using the cosine similarity between the pseudo-word vectors of story $d_i$ and $d_j$. Note that these pseudo-word vectors are weighted by term frequency-inverse document frequency (TF-IDF) [30], a popular method for exploring varying importance of different words, and normalized into unit vectors. Note that the word selection and grouping processes are completely unsupervised – there are no supervised annotations (e.g., topic labels or search relevance) provided.

### 4.3.2 Visual near-duplicate detection

As shown in Fig. 1, near-duplicates often occur in stories of the same topic. Detection of near-duplicates provides great potential for constructing contextual similarities between video stories. In [15], we also found that using near-duplicate alone, 40%-65% of stories from the same topic can be detected with almost zero false alarm rate.

For automatic detection of near-duplicates, we adopted the parts-based statistical model developed in [31]. First, salient parts are extracted from an image to form an attributed relational graph (ARG). Given two candidate images, detection of near-duplicate pairs is formulated as a hypothesis testing problem and solved by modeling the parts association between the corresponding ARGs, and computing the posterior probability (whether the image pair are near-duplicates). The detection score can then be used to measure the near-duplicate similarity between two images, or thresholded to make a binary decision.

The parts-based duplicate scores are defined between key-frame pairs (i.e., one key-frame per video shot). We represent the story-level similarity in visual duplicates as $\psi_v(d_i, d_j)$, which takes the highest duplicate scores between key-frames of story $d_i$ and $d_j$ respectively since there are usually multiple keyframes within a story (cf. Fig. 1-(d)). Note that the duplicate similarity is normalized to $[0, 1]$ by a sigmoid function.

### 4.3.3 Multimodal fusion

We use a linear weighted fusion method for combining the story-level text and visual duplicate similarities. Such a linear fusion model, though simple, has been shown to be adequate to fuse visual and text modalities in video retrieval and concept detection [6, 4]. For stories $d_i$ and $d_j$, the fused story-level similarity is

$$s_{ij} = s(d_i, d_j) = w_t \cdot \psi_t(d_i, d_j) + (1 - w_t) \cdot \psi_v(d_i, d_j), \quad (5)$$

where $w_t \in [0, 1]$ is the weight on the story-level text similarity $\psi_t$ between each story pair and $\psi_v$ is the visual similarity. The linear weights $w_t$ are later tested experimentally and will be further discussed in section 5.4.1.

# 5. RESULTS AND DISCUSSIONS

## 5.1 Experiments of variant strategies for fusing and reranking

The performances of variant methods proposed in section 3.3 are shown in Table 1. Apparently, FR method, which considers only the recurrent frequency over 2580 automatically detected stories has the worst performance among the methods. It is understandable since the recurrence frequency does not necessarily match the search topics. However, it is interesting to see that frequency alone produces a non-trival search results with MAP at .109, which is much better than .048 MAP[6] if results are returned simply based on random ranking. This is surprising since the FR result is completely independent of the query; it returns the same ranked list of stories regardless of the query. We hypothesize that the stories related to more frquent topics are more likely to be relevant to users' queries. This is intuitive because news producers (or the moderators of TRECVID search benchmark) are most likely to report ( or pick up) the topics that viewers will find interesting and important.

Nevertheless, in the FRTS method, averaging the frequency-based ranking and the initial text search scores does show noticeable performance gains – 16.5% and 10.2% over two constituent text search tools ("text-okapi" and "text-svm") respectively. It confirms that the recurrence frequency, measured by the stationary probability over the multimodal context graph, is an adequate criterion for ranking the story search. However, when the text scores are used as prior in the random walk process, the FRTP method has almost no improvement over the initial text search results. This could be due to the context graph, when including all the video stories, is too large to ensure the convergence of the random walk process to the desired semantic target.

Clearly the partial ranking – context ranking over a subset stories with positive text-based search relevance scores – outperforms the methods using the full set of stories. It might be that the search task is still dominated by the text modality. Filtering with text search scores to obtain a subset of relevant stories guarantees the performance of adding context ranking no worse than the baseline performance.

Similar to the full reranking case discussed above, considering only recurrence frequency on those relevant stories does not bring any gains, as shown in Table 1. However, in PRTS, averaging the context ranking score $\mathbf{x}_\pi$ and the initial text search scores does improve the average performance over the 24 queries. The most significant improvement comes from the PRTP method, where the personalization vector $\mathbf{v}$ is used as the random walk prior which enforces the random walk process not only going through the multimodal similarity links but also taking into account the initial text search scores (cf. Eqn. 2). The average relative improvement over all queries are significant – 32.5% and 21.6% over the two text-search baselines using "text-okapi" and "text-svm" respectively. Furthermore, the performance improvements are consistent – almost all queries are improved, as shown in Figures 5 and 6. Our conjecture is that the random walk framework utilizes the recurrent stories across sources, while the initial text scores are used as a prior for the personalization vector $\mathbf{v}$ in the random walk process. More experiment breakdowns are explained in section 5.2.

[6]In TRECVID 2005, the random-guess (story-level) APs for most queries results are in poor performance with AP much lower than .048. But there are four dominant ones, most of which correspond to the queries for commonly-observed objects, such as "people_banner" (161) with AP=.144, "meeting_table" (163) with AP=.123, "road_car" (168) with AP= .159, and "office" (172) with AP=.140.

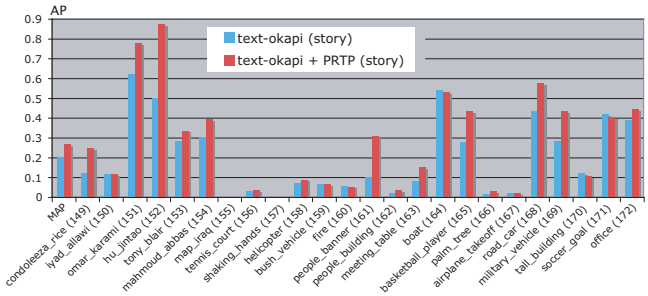


**Figure 5: Performance (in story-level AP) of PRTP at depth 20 across topics based on the text-based search set "text-okapi." The relative MAP improvement over all queries is 32.45% and that over named-people queries (topic 149-154) is 40.7%.**

In Table 1, the PRTP-KNN method, which selects the $K$ most relevant neighbors for each node in the context graph, does not improve the performance in both text-based search sets. Actually, the performances degrade when $K$ is small (below 100 on the average). This is interesting since it implies that the required connectivity, in order to exploit the power of the context graph, cannot be too small for this application domain. However, reducing the neighborhood size is still desirable since it translates into less computation load in executing the iterative random walk procedure.

In the following experiments, we have extensive experiments based on the PRTP method to see how context reranking works for different search topics, and how its performance is affected by the settings of multimodal weights in defining document similarity and reranking criteria.

## 5.2 Analysis of reranking performance over different search topics

The breakdowns of the PRTP performance across topics using the two text search methods, "text-okapi" and "text-svm," are shown in Fig. 5 and 6 respectively. The overall (story-level) MAP improvement in "text-okapi" is from 0.204 to 0.271 (32.5%) and 0.274 to 0.333 (21.6%) in "text-svm." The former is larger since it initially has a lower performance and has more space for improvement. More interestingly, the relative improvements in the people-related queries (topic 149 to 154) are significant in both sets and are 40.7% and 19.1% respectively. The improvement is larger for people-based queries than general queries since these topics are often covered by multiple news sources. Even with poor

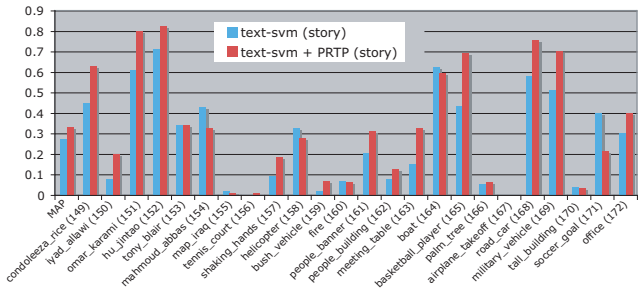


**Figure 6: Performance (in story-level AP) of PRTP at depth 20 across topics based on text-base search set "text-svm." The relative MAP improvement over all queries is 21.6% and that over named-people queries (topic 149-154) is 19.1%.**

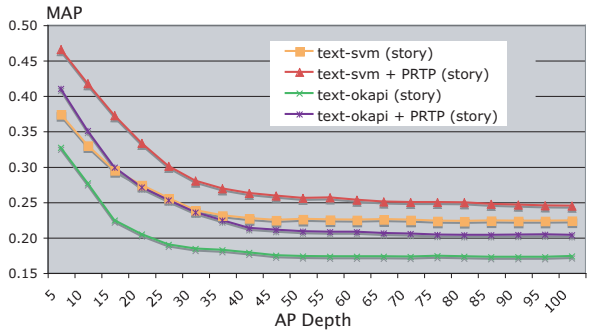


**Figure 7: Consistent performance improvements for PRTP method evaluated at variant depths in both text search sets.**

ASR/MT, the use of recurrent patterns, especially the visual duplicates, greatly improves the final search performance.

The PRTP method improves almost all queries with many queries showing significant gains, as depicted in Fig. 5 and 6. Even for the few queries that did not benefit, none has significant loss. Besides those people queries, major improvements also come from queries such as "basketball_player" (165). It was related to an important topic of "NBA brawl," in which the news about the basketball players fighting with the fans are covered across channels. So does the query "military_vehicle" (169), which consists largely of Iraq-related news stories. Another one is "people_banner" (161), though it includes no specific objects, it is mentioned in a few news events (from different topics) covered by multiple sources. Because of these cross-source relations, the inclusion of contextual relationships proves to be beneficial.

Compared to the traditional techniques of text-based query expansion, our proposed method is advantageous since it does not require explicit uses of new words or visual models which are often needed in expanded queries. In contrast, our method leverages the underlying (multimodal) similarity among documents and uses them to establish context and refine the initial search scores.

The previous discussions are based on the AP at depth[7] 20, which is a reasonable number of results that users might like to browse in typical search scenarios. In such case, an improvement from AP 0.50 to 0.87 is significant, as in the query "hu_jintao" (152) shown in Fig. 5. Nevertheless, we are also interested in the performance at different return depths. The results are plotted in Fig. 7, which shows clear MAP improvements over text search baselines. The relative improvement decreases as the return depth increases since the results are gradually dominated by irrelevant stories[8].

## 5.3 Time complexity for reranking

A single iteration of the Power Method (cf. section 3.2) consists of the single matrix-vector multiply $\mathbf{Ax}_{(k-1)}$. Generally, this is an $O(n^2)$ operation. In our experiment, the methods converges within few iterations ranging from 3 to 8. Intuitively, the partial ranking strategies incur less computation time. Empirically, for the square matrix $\mathbf{A}_{n\times n}$ ($n$ is about 200 to 400 for partial ranking (PR) method), the Power Method implemented in *MATLAB* takes less than a second in a regular desktop with an Intel Core Duo CPU. For much large-scale databases, there are several advances in the implementation, such as the method proposed in [18] and those methods discussed in [19].

## 5.4 Parameter Sensibility

### 5.4.1 *Visual Modality Significance*

The significance of the contributions from the visual modality has been of great interest to researchers. The context graph of the random walk process includes linearly weighted combination of contributions from visual duplicates and text similarity in defining the link strength (i.e. cross-node similarity). To analyze the impact of each individual modality, we compare the performance with different text weights ranging from 0.0 to 1.0 with an increasing step of 0.05 and plot the results in Fig. 8. Interestingly, we discover that the best weight for text is 0.15 in both "text-svm" and "text-okapi." The dominant weight (0.85) for the visual modality is consistent with what we have anticipated. Our conjecture is that the initial text search has already used the text information in computing the relevance scores in response to the query. Therefore, additional gain by measuring cross-document similarity will probably come from the visual aspect (e.g., via visual duplicate detection), rather than the text modality.

Other interesting observations are related to the extreme cases of visual only and text only, shown as the two end points of each curve in Fig. 8. The performances are almost the same as the text-base search sets if the context graph considers text modality solely (i.e., $w_t$, text weight in Eqn. 5, is 1). However, when using the context graph in visual duplicates only (i.e., $w_t = 0$), the PRTP performance still achieves significant improvement (+26.1%) in the "text-okapi" and slightly (+6.5%) in "text-svm" text-based search sets. It confirms the significant contribution of visual similarity relationship in context ranking.

[7]The number of top ranking documents (stories) to be included in the search result evaluation [1].

[8]In TRECVID, the shot-level AP for video search is evaluated at depth 1000. We do not evaluate in the same way since we are looking at the story-level search performance, while TRECVID was evaluating the shot-level search accuracy. Since a typical story contains multiple shots, there are much less stories than shots relevant to each query. It is reasonable to use a smaller depth when evaluating the story search results.

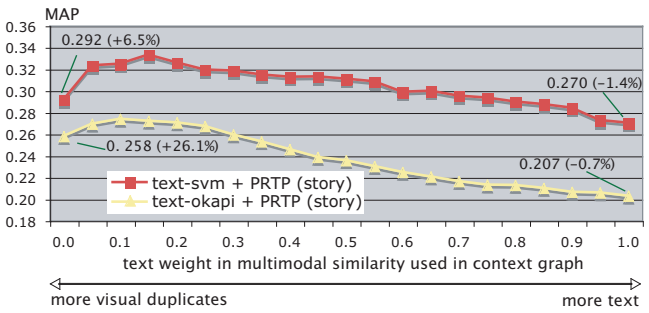


**Figure 8: Context reranking (with PRTP) with variant text similarity weights ($w_t$ in Eqn. 5) for the context graph. The numbers in the parentheses are the relative improvements from the text search result (in MAP). The best performance of the ratio of text vs. duplicates is around .15:.85.**

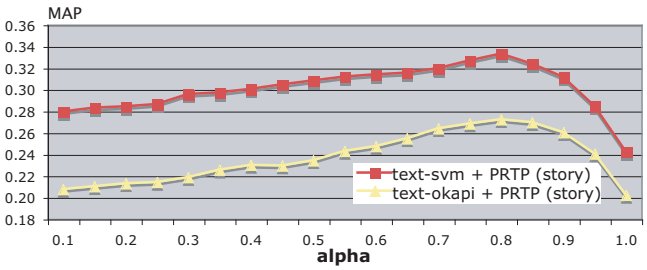


**Figure 9: PRTP with variant $\alpha$ in Eqn. 2. The best performance in each set is when $\alpha = 0.8$. The text weight $w_t$ in Eqn. 5 is set to 0.15.**

### 5.4.2 $\alpha$ Sensibility

The use of $\alpha$ in Eqn. 1 plays an important role in balancing the personalization vector $\mathbf{v}$ (i.e., the text search prior in this experiment) and the multimodal contextual relationships walk process. The value of $\alpha$ also impacts the performance and convergence speed [18, 24]. With $\alpha$ close to 1, the random walk process relies almost entirely on the multimodal contextual similarities and ignores the text search prior; hence, the performance degrades sharply. Based on our empirical comparisons, the best setup for both data sets is $\alpha = 0.8$, which is close to but slightly less than what have been reported in large-scale web applications. The best value reported in [18] was around 0.9 and 0.85 in [19]. The authors of [24] reported that the performance reaches a plateau as $\alpha$ grows from 0.5 to 0.9. The above findings are important. Although the document links in our context graph are constructed based on implicit relations (e.g., visual near-duplicates) instead of explicit document links, such implicit links are proved to play a significant role in multimedia document reranking, in a way similar to that played by the page links in web document ranking.

## 6. OTHER RELATED WORKS

There are a few works in the video research community utilizing the random walk framework. Authors of [20] formulate the *normalized cut* problem [25] in video segmentation as a random walk process. By normalizing the similarity matrix of pair-wise pixels, one obtains the stochastic matrix and then utilizes the second largest eigenvector or sets of eigenvectors [22] to do image segmentation. It differs from our approach as it is used for clustering of pixels; we focus on the high-level contextual similarities between stories; meanwhile, the text modality in our approach is used as a prior in the random walk process. We are primarily interested in the ranking of the stationary probability rather than the pixel similarities in the spectral (eigenvector) space.

The authors of [24] utilized the random framework to associate image regions to certain keywords by constructing a graph with nodes composed of annotated keywords, images, and their regions. The links between the nodes are binary by thresholding similarities between nodes or the existence of annotated words to certain images.

In addition, He *et al.* proposed an approach called *ImageRank* in [13] to selecte representative images from a collection of images. The selection criterion is determined by the stationary probability derived from the (normalized) transition matrix modeled by low-level features (i.e., color correlogram) between images. The approach does not address the multimodal search problem. In addition, only the image-level similarity is considered, unlike the combined text-visual similarity between video stories.

Our reranking processes are based on the initial text search results. Such methods utilize the initial search scores as some type of pseudo supervision in order to improve the search accuracy. Conceptually, they are related to approaches such as Transductive Learning [17] and Co-training [3]. These two paradigms consider the problem of using a large number of unlabeled samples to boost the performance of a learning algorithm when only a small set of labeled examples are available. However, in the search reranking problem, we cannot exactly locate what positive data might be in the initial text search results. Instead, only the (noisy) search relevance scores from the initial text search are available. The use of such text search scores in reranking multimedia documents has not been explored before ,except the work done in [16], which however performs reranking at the shot level without considering the multimodal document similarity and the powerful context graph reranking approach.

## 7. CONTRIBUTIONS AND CONCLUSIONS

We proposed a novel and effective reranking process for video search, which requires no search examples or highly-tuned complex models. The approach utilizes the recurrent patterns commonly observed in large-scale distributed video databases and employs the multimodal similarity between video stories to improve the initial text search results.

The unique contributions of the paper include the use of the multimodal similarities to define inter-video document contextual relationships and the solution of the reranking process via random walk over the context graph. The extensive experiments over the large TRECVID video benchmark (170 hours) confirm the effectiveness of the proposed method – 32% average performance gain. Our sensitivity study also verifies the large impact of the visual modality in combining text and visual features in the reranking process.

The proposed contextual reranking framework is general and may be applied to multimedia document retrieval in other domains that have strong contextual links among information from multiple sources. For example, media-rich

social networks or blog spaces will be promising areas for applying the proposed methods.

Another direction of future work is to incorporate the large number of semantic concept detectors available in [29] to extract high-level semantic descriptions of visual content, and then use them in expanding the contextual relations between videos. It will be interesting to analyze the impact of such large-scale high-level description, in comparison with the text features and visual duplicate used in this paper.

## 8. ACKNOWLEDGMENT

This material is based upon work funded in whole by the U.S. Government. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not neces- sarily reflect the views of the U.S. Government.

## 9. REFERENCES

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

[2] J. Battelle. *The Search: How Google and Its Rivals Rewrote the Rules of Business and Transformed Our Culture*. Portfolio Trade, 2006.

[3] A. Blum and et al. Combining labeled and unlabeled data with co-training. In *Annual Workshop on Computational Learning Theory*, 1998.

[4] M. Campbell and et al. IBM Research TRECVID-2006 Video Retrieval System. In *NIST TRECVID workshop*, Gaithersburg, MD, November 2006.

[5] J. G. Carbonell and et al. Translingual information retrieval: A comparative evaluation. In *International Joint Conference on Artificial Intelligence*, 1997.

[6] S.-F. Chang and et al. Columbia University TRECVID-2006 video search and high-level feature extraction. In *TRECVID Workshop*, Waishington DC, 2006.

[7] T.-S. Chua and et. al. TRECVID 2004 search and feature extraction task by NUS PRIS. In *TRECVID Workshop*, Waishington DC, 2004.

[8] A. K. Dey. Understanding and using context. *Personal and Ubiquitous Computing*, 5(1), February 2001.

[9] K. M. Donald and A. F. Smeaton. A comparison of score, rank and probability-based fusion methods for video shot retrieval. In *International Conference on Content-Based Image and Video Retrieval (CIVR)*, Singapore, 2005.

[10] R. Fergus, P. Perona, and A. Zisserman. A visual category filter for google images. In *Proceedings of the 8th European Conference on Computer Vision, Prague, Czech Republic*, May 2004.

[11] A. G. Hauptmann and M. G. Christel. Successful approaches in the TREC Video Retrieval Evaluations. In *ACM Multimedia 2004*, New York, 2004.

[12] A. G. Hauptmann and et al. Multi-Lingual Broadcast News Retrieval. In *NIST TRECVID workshop*, Gaithersburg, MD, November 2006.

[13] X. He and et al. Imagerank : spectral techniques for structural analysis of image database. In *ICME*, 2003.

[14] W. Hsu, L. Kennedy, S.-F. Chang, M. Franz, and J. Smith. Columbia-IBM news video story segmentation in trecvid 2004. Technical Report ADVENT #207-2005-3, Columbia Universiry, 2005.

[15] W. H. Hsu and S.-F. Chang. Topic tracking across broadcast news videos with visual duplicates and semantic concepts. In *International Conference on Image Processing (ICIP)*, Atlanta, GA, USA, 2006.

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

[17] T. Joachims. Transductive inference for text classification using support vector machines. In *16th International Conference on Machine Learning*, pages 200–209, 1999.

[18] S. D. Kamvar, T. H. Haveliwala, C. D. Manning, and G. H. Golub. Extrapolation methods for accelerating pagerank computations. In *International WWW Conference*, Budapest, Hungary, 2003.

[19] A. N. Langville and C. D. Meyer. A survey of eigenvector methods for web information retrieval. *SIAM Review*, 47(1):135–161, 2005.

[20] M. Meila and J. Shi. Learning segmentation with random walk. In *Neural Information Processing Systems Conference (NIPS)*, pages 873–879, 2001.

[21] A. Natsev and et al. Learning the semantics of multimedia queries and concepts from a small number of examples. In *ACM Multimedia*, pages 598–607, Singapore, 2005.

[22] A. Y. Ng and et al. On spectral clustering: Analysis and an algorithm. In *Neural Information Processing Systems Conference (NIPS)*, 2002.

[23] L. Page and et al. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford Digital Library Technologies Project, 1998.

[24] J.-Y. Pan and et al. Gcap: Graph-based automatic image captioning. In *International Workshop on Multimedia Data and Document Engineering*, Washington, DC, USA, 2004.

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

[26] N. Slonim and N. Tishby. Document clustering using word clusters via the information bottleneck method. In *SIGIR*, Athens, Greece, 2000.

[27] C. G. M. Snoek and et. al. The MediaMill TRECVID 2006 Semantic Video Search Engine. In *NIST TRECVID workshop*, Gaithersburg, MD, Nov. 2006.

[28] R. Yan, A. Hauptmann, and R. Jin. Multimedia search with pseudo-relevance feedback. In *International Conference on Image and Video Retrieval*, Urbana-Champaign, IL, USA, 2003.

[29] A. Yanagawa and et al. Columbia university's baseline detectors for 374 LSCOM semantic visual concepts. Technical Report ADVENT #222-2006-8, Columbia Universiry, 2007.

[30] Y. Yang and et al. Learning approaches for detecting and tracking news events. *IEEE Intelligent Systems*, 14(4), 1999.

[31] D.-Q. Zhang and et al. Detecting image near-duplicate by stochastic attributed relational graph matching with learning. In *ACM Multimedia*, New York, 2004.