Reranking Methods for Visual Search

Winston H. Hsu
National Taiwan University

Lyndon S. Kennedy and Shih-Fu Chang
Columbia University

Most semantic video search methods use text-keyword queries or example video clips and images. But such methods have limitations. Here the authors introduce two reranking processes for image and video search that automatically reorder results from initial text-only searches based on visual features and content similarity.

The continuing growth of online video data, digital photos, and 24-hour broadcast news has helped make video and image retrieval an active research area. Current semantic video search usually builds on text search by using text associated with video content, such as speech transcripts, closed captions, or video optical character recognition (OCR) text. Adding multiple modalities such as image features, audio, face detection, and high-level concept detection can improve text-based video search systems.[1-4] Researchers often achieve this using multiple query example images, applying specific concept detectors, or incorporating highly tuned retrieval models for specific query types. However, because it's difficult to acquire the extra image information or build the models in advance, such approaches are impractical for large-scale applications.[5]

It's also difficult for users to acquire example images for example-based queries. When users provide image examples, the examples are often few in number and fail to capture the relevant regions of high-dimensional feature space that might contain the true positive videos. Practically, users prefer "search by text" rather than "search by examples."[6]

To address the problems of example-based video search approaches and avoid the use of specialized models, we conduct semantic video searches using a reranking method that automatically reorders the initial text search results based on visual cues and associated context. We developed two general reranking methods that explore the recurrent visual patterns in many contexts, such as the returned images or video shots from initial text queries, and video stories from multiple channels.

Visual reranking

Assume our initial text search returns n visual documents $\{d_1, d_2, \ldots, d_n\}$, as Figure 1 illustrates. These documents might include Web pages, images, and video stories. The visual reranking process improves search accuracy by reordering the visual documents based on multimodal cues extracted from the initial text search results and any available auxiliary knowledge. The auxiliary knowledge can be features extracted from each visual document or the multimodal similarities between them.

Pseudorelevance feedback[2,7-9] is an effective tool for improving initial text search results in both text and video retrieval. PRF assumes that a significant fraction of top-ranked documents are relevant and uses them to build a model for reranking the search result set.[7] This is in contrast to relevance feedback, in which users explicitly provide feedback by labeling results as positive or negative.

Some researchers have implemented the PRF concept in video retrieval. Chua et al., for example, used the textual information in top-ranking shots to obtain additional keywords to perform retrieval and rerank the baseline shot lists.[2] Their experiment improved mean average precision (MAP) from 0.120 to 0.124 in the TRECVID 2004 video search task (see http://www-nlpir.nist.gov/projects/trecvid; MAP is a search performance metric used in TRECVID, a National Institute of Standards and Technology-sponsored conference focused on information retrieval in digital video). Other researchers sampled the pseudonegative images from the lowest rank of the initial query results, used the query videos and images as the positive examples, and formulated retrieval as a classification problem.[8] This improved the search performance from MAP 0.105 to 0.112 in TRECVID 2003.

Reranking approaches

Patterns with strong visual similarities often recur across diverse video sources. Such recurrent images or videos frequently appear in image search engines (such as Yahoo or Google) and photo-sharing sites (such as Flickr). In earlier work,[10] we analyzed the frequency of such recurrent patterns (in terms of visual duplicates) for

1070-986X/07/$25.00 © 2007 IEEE Published by the IEEE Computer Society

cross-language topic tracking (a large percentage of international news videos share common video clips, near duplicates, or high-level semantics). Based on our observations, we propose two general, visual search reranking approaches.

We first propose an image search reranking approach called *information bottleneck (IB)* reranking.[5,11] This approach explores the fact that in image search, multiple similar images are often spread across different spots in the top-level pages of the initial text search results.[5] IB reranking revises the search relevance scores to favor images that occur multiple times with high visual similarity and high initial text retrieval scores. Visual similarity is typically based on matching image features extracted from image frames in video shots. Thus, we can easily apply it to rerank search results at the image or video shot levels.

The preferred search level is sometimes higher than the image or shot level—that is, a semantic level such as news stories or scenes in films. In such cases, feature-based similarity between visual documents isn't well defined. To address this issue, we propose *contextual links*. For example, two stories share the same or similar contexts if they're related to the same topic and thus share some common or near-duplicate visual footage. Our *context-reranking* approach explores such contextual similarity among high-level video stories using a random walk framework.[12]

IB-reranking: Image/video reranking based on recurrent frequency

In image/video search systems, we're given a query or a statement of information need, and must estimate the relevance $R(x)$ of each image or video shot in the search set, $x \in X$, and order them by relevance scores. Researchers have tested many approaches in recent years—from simply associating video shots with text search scores to fusing multiple modalities. Some approaches rely on user-provided query images as positive examples to train a supervised classifier to approximate the posterior probability $P(Y|X)$, where Y is a random variable representing search relevance.[9] They then use the posterior probability for $R(x)$ in visual ranking.

Text search results generally put certain relevant images near the top of the return set. In large image or video databases, in particular, some positive images might share great visual similarity, but receive quite different initial text search scores. For example, Figure 2b (next page) shows the top 24 shots from a story-level text search for TRECVID 2005 topic 153, "Find shots of Tony Blair." We can see recurrent relevant shots of various appearances among the return set, but they're mixed with some irrelevant shots (for example, anchor or graphic shots).

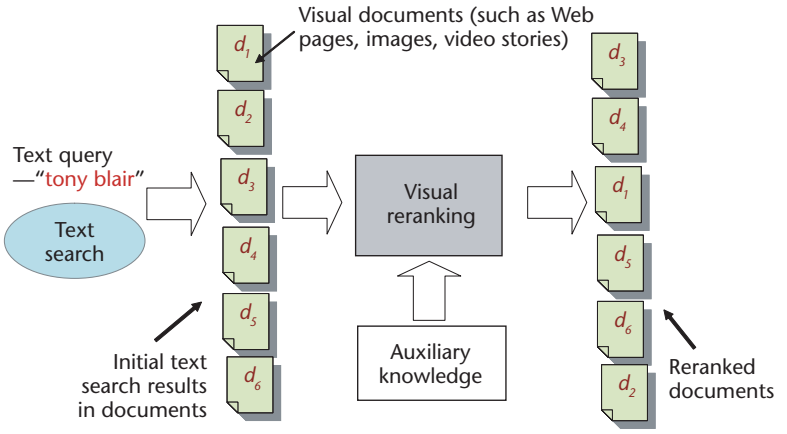

Figure 1. The proposed visual reranking process improves search accuracy by reranking visual documents (such as Web documents, images, and videos) based on multimodal cues extracted from initial text search results.

Thus, instead of user-provided search examples, we consider the search posterior probability, estimated from initial text-based results in an unsupervised fashion, and the recurrent patterns (or feature density) among image features, and use them to rerank the search results. A straightforward implementation of the idea is to fuse both measures—search posterior probability and feature density—in a linear fashion. This fusion approach is commonly used in multimodal video search systems.[5] We can formulate the approach as

$$R(x) = \alpha p(y|x) + \beta p(x) \quad (1)$$

where $p(y|x)$ is the search posterior probability, $p(x)$ is the retrieved image's feature density, and α and β are linear fusion scalars.

Equation 1 incurs two main problems, which our prior experiment confirms.[5] The posterior probability $p(y|x)$, estimated from the text query results and (soft) pseudolabeling strategies, is noisy; we need a denoised representation for the posterior probability. In addition, the feature density estimation $p(x)$ in Equation 1 might be problematic because recurrent patterns that are irrelevant to the search (for example, anchors, commercials, and crowd scenes) can be frequent. Instead, we should consider only those recurrent patterns within buckets (or clusters) of higher relevance. To exploit both search relevance and recurrent patterns, we represent the search relevance score $R(x)$ as

$$R(x) = \alpha p(y|c) + \beta p(x|c) \quad (2)$$

Search topic:
"Find shots of Tony Blair"
and search examples

(a)

(b) Text search results

(c) IB reranked results + text search

Figure 2. (a) TRECVID 2005 search topic 153, "Find shots of Tony Blair" and four search examples; (b) top 24 returned shots from story-level text search (0.358 average precision) with query terms "tony blair," and (c) top 24 shots of information bottleneck (IB) reranked results (0.472 AP) with low-level color and texture features. The red triangles signal true positives.

where $p(y|c)$ is a denoised posterior probability smoothed over a relevance-consistent cluster c, which covers image x; and $p(x|c)$ is the local feature density estimated at feature x. The cluster-denoising process has been effective in text search.[13] Meanwhile, we use local feature density $p(x|c)$ to favor images occurring multiple times with high visual similarity. The choice of parameter α or β will affect the reranked results. In our preliminary experiments, the denoised posterior probability $p(y|c)$ was more effective and played a more important role in search relevance than the pattern recurrence within the same relevant clusters. Accordingly, an intuitive approach is to let α be significantly larger than β so that the reranking process first orders clusters at a coarse level and then refines the image order in each cluster according to local feature density.

Two main issues arise in this proposed approach:

- How can we find the relevance-consistent clusters, in an unsupervised fashion, from noisy text search results and high-dimensional visual features?
- How can we use the recurrent patterns across video sources?

To address the first problem, we adopt the IB principle,[5,11] which finds the optimal clustering of images that preserves the maximal mutual information about the search relevance. We iteratively update the denoised posterior probabilities $p(y|c)$ during the clustering process. We then estimate the feature densities $p(x|c)$ from each cluster c accordingly.

Figure 3 illustrates this idea. In the figure, the reranking method discovers four relevance-consistent clusters automatically. Images of the same cluster (that is, C_1) have the same denoised posterior probability $p(y|c)$, but might have recurrent patterns with different appearances. For example, C_1 has three regions with high density in the feature space. We first rank the image clusters by posterior probability $p(y|c)$ and then order within-cluster images by the local feature density $p(x|c)$. In short, the visually consistent images that occur most frequently within higher-relevance clusters will receive higher ranks. Note that the visual features we adopted are grid color moments and Gabor texture. More technical details are available elsewhere.[5]

Context reranking: Story reranking based on multimodal similarity

Our context-reranking method uses the multimodal similarity between video stories to improve the initial text query results. The example in Figure 4 illustrates the motivation for this approach. An initial text query retrieves video stories with the keywords "tony blair"; however, it doesn't retrieve certain relevant stories because of the lack of keyword annotations associated with such videos. We can use contextual links (such as visual duplicates and text) to link such missing stories to the initial text queries and further improve the search accuracy.

Because stories don't have explicit links, we formulate our approach as a random walk over a

graph whose nodes represent stories in the search set. These nodes are connected by the edges, weighted with pairwise multimodal contextual similarities (for example, visual duplicates and text tokens).[10] We use the random walk's stationary probability to compute stories' final scores after reranking. The random walk is biased toward stories with higher initial text search scores—a principled way to combine initial text search results and stories' implicit contextual relationships.

Other researchers in the video research community have used the random walk framework. Meila and Shi formulate the normalized cut problem in video segmentation as a random walk process.[14] By normalizing the similarity matrix of pairwise pixels, they obtain a stochastic matrix and use the second largest eigenvector or sets of eigenvectors to perform image segmentation. Their approach differs from ours in that it's used for pixel clustering, whereas we focus on the high-level contextual similarities between stories. Meanwhile, our approach uses the text modality as a prior in the random walk process. Moreover, we're primarily interested in ranking the stationary probability rather than the similarities in the spectral (eigenvector) space.

Video story similarity. To benefit from the multiple modalities, the story-level similarity between stories consists of visual duplicates and text similarities (in terms of cue word clusters), which are linearly fused.[10]

We represent each story's text modality by compact 120D cue word clusters or pseudowords,[10,11] where words in the same cue word cluster are associated with the same semantic. We then compute the story-level similarity using the cosine similarity between the pseudoword vectors of story pairs. We also weight these pseudoword vectors by term frequency–inverse document frequency (TF-IDF)[15] and normalize them into unit vectors.

As we describe elsewhere,[10] visual duplicates are effective in topic threading. This approach represents images in a statistical parts-based model. Given two candidate images, we formulate near-duplicate detection as a hypothesis-testing problem and solve the problem by modeling the parts association between images. We then represent the story-level similarity in visual duplicates and take the highest duplicate scores between story pair key frames. We normalize the duplicate similarity to [0,1] using a sigmoid function.

Random walk on visual story-level similarities. In the random walk framework over the

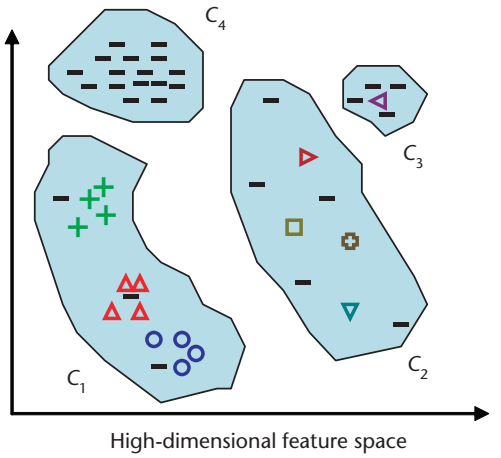

Figure 3. Four search relevance-consistent clusters, C_1 to C_4, where C_1 has the highest denoised posterior probability $p(y = 1|c)$ and C_4 has the lowest. The symbol "–" is a pseudonegative, and the others are pseudopositives. Pseudopositives in the same shape are assumed to have a similar appearance. (Note that the "hard" pseudolabels are for illustration only; we use "soft" labels in this work.)

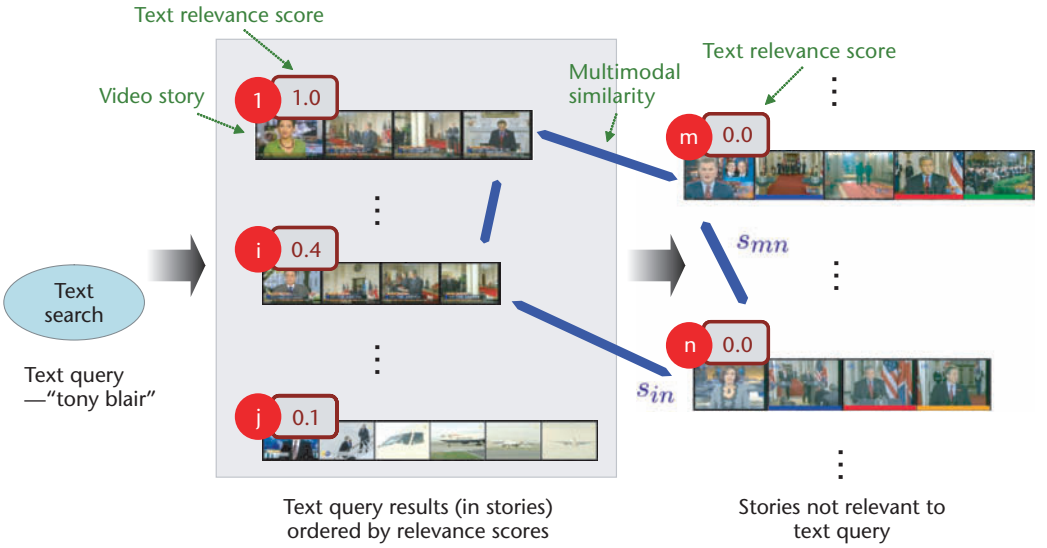

Figure 4. A video search that benefits from multimodal story-level similarity in a large-scale video database, even with unreliable automatic spoken language/machine translation (ASR/MT) transcripts. Though not relevant to the text query, certain stories can be boosted because of their closeness to some relevant text query stories by the multimodal similarity (shown in the right panel) consisting of text, visual duplicates, and high-level concepts.

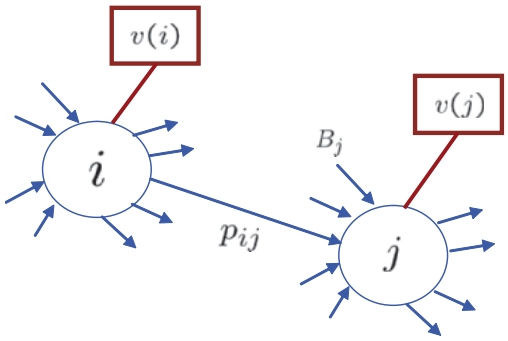

Figure 5. Context graph for random walk over nodes (stories). In the graph, **i** *and* **j** *are node indexes with their original text search scores* **v**(**i**) *and* **v**(**j**)*;* $\mathbf{p}_{ij}$ *is the transition probability from node* **i** *to* **j***; and* $\mathbf{B}_j$ *are edges back to node* **j**.

context graph (the "auxiliary knowledge" in Figure 1), stories are nodes and the edges between them are weighted by multimodal contextual similarities, as Figure 5 illustrates. We use the context-ranking score—the stationary probability of the random walk over the context graph—to represent the search relevance scores.

Assume 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 use a stochastic (transition) matrix

$$\mathbf{P} \equiv \left[p_{ij}\right]_{n\times n} = \left[p\left(j|i\right)\right]_{n\times n}$$

to govern a random walk process's transition. In the matrix, 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 that instance. The $n \times 1$ stationary probability $\mathbf{x}_\pi \equiv \lim_{k\to\infty} \mathbf{x}_{(k)}$ is the random walk process's state probability as the time instance proceeds to infinity, if the convergence conditions are satisfied.

In this context, we consider both the multimodal similarities, or transition probabilities **P**, between stories and the original (normalized) text search scores **v**, or *personalization vector*. In this framework, the state probability $x_{(k)}(j)$ of node j at time instance k is

$$x_{(k)}\left(j\right) = \alpha \sum_{i\in B_j} x_{(k-1)}\left(i\right) p_{ij} + \left(1-\alpha\right) v\left(j\right) \tag{3}$$

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

Equation 3 is an interesting interpretation of the random walk process illustrated in Figures 4 and 5. Intuitively, $x_{(k)}(j)$ is parameterized by its neighboring nodes B_j at time instance $k-1$ and its own initial text scores $v(j)$. We then linearly fuse both terms with weights α and $1-\alpha$, respectively. For the first term in Equation 3, we consider not only the state probabilities of its neighbors B_j but also their corresponding transition probabilities—that is, 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 node j's neighbors and its initial text scores. Multimodal search and concept detection research often use linear fusion, and have shown it to be effective.[5]

We then update the relationship in Equation 3 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 Equation 3, node j's stationary probability is

$$x_\pi\left(j\right) = \alpha \sum_{i\in B_j} x_\pi\left(i\right) p_{ij} + \left(1-\alpha\right) v_j \tag{4}$$

Naturally, we have $\Sigma_j x_\pi(j) = 1$ and $\mathbf{e} = [1]_{n\times 1}$ is an n-dimensional column vector with all 1. If we assume $\mathbf{E} = \mathbf{e}\mathbf{v}^T$ and go through some algebra, we get the following:

$$\begin{aligned} \mathbf{x}_\pi^T \equiv \left[x_\pi\left(j\right)\right]_{1\times n} &= \alpha \mathbf{x}_\pi^T \mathbf{P} + \left(1-\alpha\right)\mathbf{x}_\pi^T \mathbf{E} \\ &= \mathbf{x}_\pi^T \left[\alpha \mathbf{P} + \left(1-\alpha\right)\mathbf{E}\right] \end{aligned} \tag{5}$$

Equation 5 has almost the same formulation as the PageRank algorithm, in which the stationary probability can be derived through an efficient algorithm, the Power Method.[16]

We need to derive the context graph transition probability from the raw story-level affinity matrix **S**. Intuitively, we conduct the normalization and ensure that each row equals 1. That is, $p_{ij} = s_{ij}/\Sigma_k s_{ik}$. A random walk process will more likely jump to the nodes with higher (contextual) similarities. Like the PageRank algorithm, we handle the dangling problem and set all zero rows of affinity matrix **S** as $\mathbf{e}^T$. The affinity matrix **S** is composed of pairwise similarities between stories. The story-level similarity consists of visual duplicates and text, as we discussed earlier. We tested variant modality weights between visual duplicates and text.

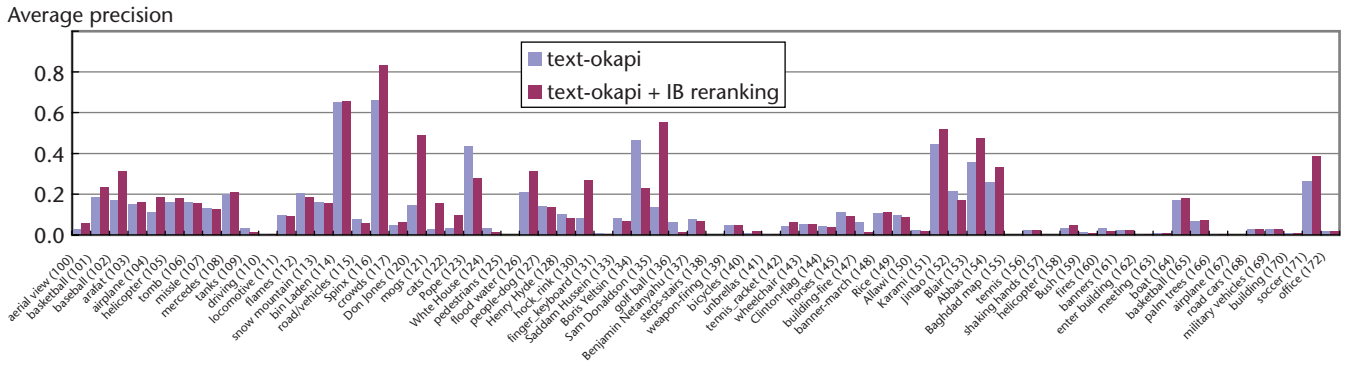

Figure 6. Performance of IB reranking and the baseline text search across all queries of TRECVID 2003–2005.

Experiments

For our experiments on the two reranking methods, we used TRECVID 2003–2005 data sets (see http://www-nlpir.nist.gov/projects/trecvid).

Data set

TRECVID 2003–2004 contains 133 and 177 hours of videos, respectively, with English ASR transcripts from 1998 for both CNN and ABC. The TRECVID 2005 data set contains 277 international broadcast news videos, which includes 171 hours of videos from six channels in three languages (Arabic, English, and Chinese). The time span is 30 October to 1 December 2004. The National Institute of Standards and Technology provided the ASR and MT transcripts.

Performance metrics. For the performance metric, we adopted noninterpolated AP, which corresponds to the area under a (noninterpolated) recall/precision curve. Because AP only shows a single query's performance, we use MAP to measure the average performance over sets of queries in a test set.

Text search and broadcast news video stories. In earlier work, we showed that the best approach to text searches over speech-recognition transcripts from international broadcast news videos is to use transcripts from the entire story.[5] This makes sense because the true semantic relationships between images and the text transcripts exist at the story level: a concept mentioned in the text is likely to appear in the video somewhere within the same story, but unlikely to appear in the next or previous story. We can automatically extract story boundaries with reasonable accuracy by analyzing the image content's visual characteristics and the anchorperson's speech.[4] If we apply the Okapi method[5] to retrieve these stories, we get the text-based search baseline, "text-okapi."

IB reranking performance

We conducted IB reranking on all TRECVID 2003–2005 queries. We found that IB reranking consistently improves the performance (MAP) of the text search baseline (that is, text-okapi) and achieves 20.8-, 17.7-, and 23.0-percent relative improvement, respectively, for each year.

Figure 6 lists the performance (text-okapi vs. text-okapi + IB reranking) in AP across all queries. As the figure shows, IB reranking improves or retains the text search baseline's performance for most queries. IB reranking is more beneficial for queries with salient recurrent patterns: "Sam Donaldson" (135), "Omar Karami" (151), "Blair" (153), "Mahmoud Abbas" (154), "baseball" (102), "Spinx" (116), "Dow Jones" (120), and "soccer" (171). This makes sense because the approach, although it requires no example images, tries to infer the recurrent patterns that are highly relevant to the search based on the initial search scores and visual density estimation. Specifically, the visual patterns present in the search results help boost the posterior probabilities of relevant data through denoising and local density-based reranking in each cluster (see Equation 2).

Figure 6 also shows several query topics for which performance degrades after IB reranking. These queries include "building-fire" (147), "Pope" (123), and "Boris Yeltsin" (134). On further examination, we found that relevant videos for such queries are either few or lack consistent visual patterns. For example, scenes of the pope are actually of different events, so they don't form consistent visual appearances.

IB reranking reorders images from the text output directly, requiring no external image examples. This is an important advancement in

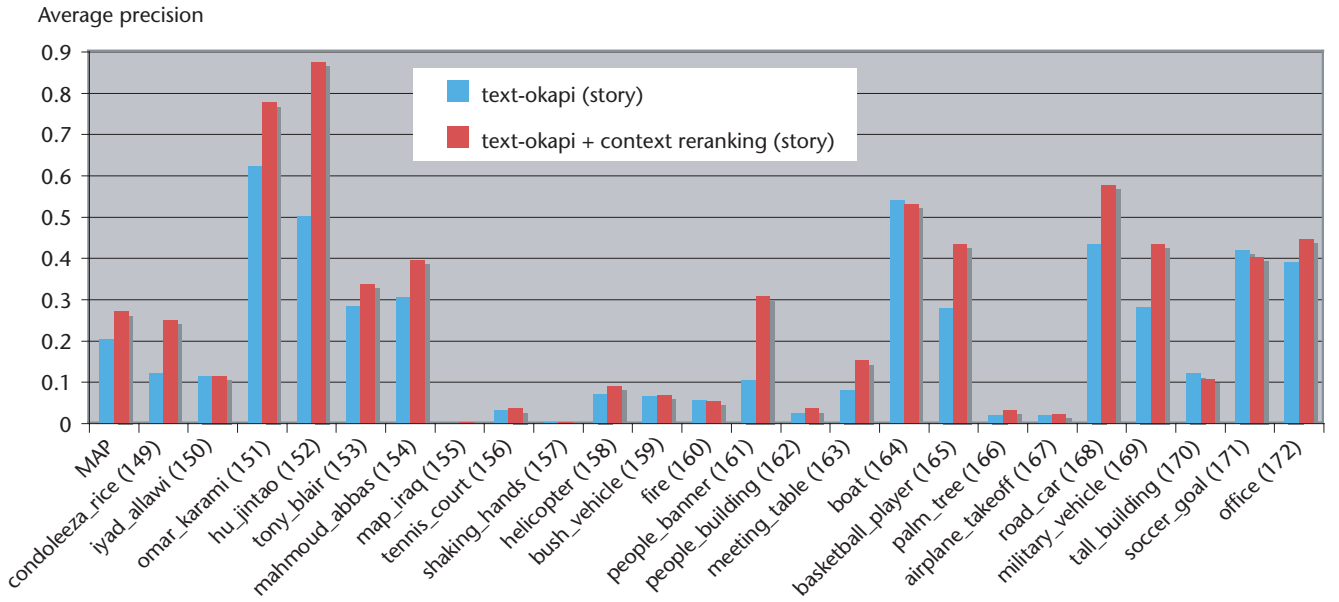

Figure 7. Performance (in story-level AP) of context reranking at depth 20 across topics based on the text-okapi text-based search set. The relative MAP improvement over all queries is 32.45 percent and over named-people queries (149–154) is 40.7 percent.

video search because users often don't have or are reluctant to provide image examples. Surprisingly, our approach is competitive with and complementary to state-of-the-art example-based search approaches.[5]

IB reranking works best when initial text search results contain a reasonable number of positive shots, and the positive shots are somewhat visually related. Such conditions often occur when the video sources are multiple concurrent channels reporting related events. This is more the case for TRECVID 2005 than for TRECVID 2003 and 2004.

In the TRECVID 2005 benchmark data, IB reranking works particularly well on queries for named persons. This is surprising, because previous work hasn't shown the visual content of example images and shots to be helpful for retrieving named persons. For example, fusing simple story-based text search with a high-performing content-based image retrieval (CBIR) system[3,9] provides only modest gains over story-based text search on the six named person queries in the TRECVID 2005 set. Story-based text search results in a MAP of 0.231 over these six queries, while fusing with the CBIR system improves MAP by 4 percent, to 0.241.

On the other hand, if we apply IB reranking after story-based text search, we get an improvement in MAP to 0.285, an improvement of more than 23 percent. So, IB reranking captures the salient visual aspects of news events contained in the search set in which particular named people appear. Such a capture is difficult using example images from sources other than the search set or from a different period of time.

Context reranking performance

Experiments with our second reranking approach used TRECVID 2005 data in which the video search ground truth was at the shot level. Because our evaluation was at the story level, we converted the shot-level ground truth into story-level ground truth. We considered a story positive if it contained at least one shot that was labeled as positive for a particular query.

Analysis of context reranking performance. Figure 7 shows the breakdowns of the context reranking performance across topics using the text-okapi text search method. The overall (story-level) MAP improvement is from 0.204 to 0.271 (32.5 percent). More interestingly, the relative improvements in the people-related queries (149 to 154) are significant, at 40.7 percent. Generally the improvement is larger for people-based queries than for general queries because the people-related topics are typically reported in major political news events worldwide. Even with poor ASR/MT, the use of recurrent patterns, especially visual duplicates, greatly improves story-ranking performance.

The context-reranking method improves almost all queries, with many queries showing significant gains, as Figure 7 shows. Even for the few queries that didn't benefit, none had significant loss. In addition to people queries, queries such as "basketball player" (165) also showed

improvement. This query was related to the topic "NBA brawl," a widely covered news item about basketball players fighting with fans. Results for the query "military vehicle" (169), which consisted largely of Iraq-related news stories, also improved.

Another query showing improvement is "people banner" (161). Although it includes no specific objects, it's covered in a few events (from different topics) in the news collections. Because of these cross-source relations, by including contextual relationships we improve precision over text-based queries.

Unlike traditional text-based query expansion techniques, our proposed method doesn't require explicit discovery of new words or visual models. In contrast, our method leverages the underlying similarities among stories and uses them to establish context and refine the initial search scores.

Significance of visual modality. The random walk process's context graph includes linearly weighted combinations of contributions from visual duplicates and text similarity. To analyze each modality's impact, we compare their performance using text weights ranging from 0.0 to 1.0 with an increasing step of 0.05, and plot the results in Figure 8. Our results show that the best weight for text is 0.15. The dominant weight (0.85) for the visual modality is consistent with our expectations. We conjecture that the initial text search used the text information in computing the relevance scores in response to the query. Therefore, additional gain by measuring cross-story similarity will probably come from the visual aspect (such as 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 endpoints of each curve in Figure 8. The performances are almost the same as the text-based search sets if the context graph considers text modality alone (that is, text weight = 1). However, when using the context graph for visual duplicates only (that is, text weight = 0), the context reranking performance still achieves significant improvement (more than 26.1 percent). This confirms the significant contribution of visual similarity in context ranking.

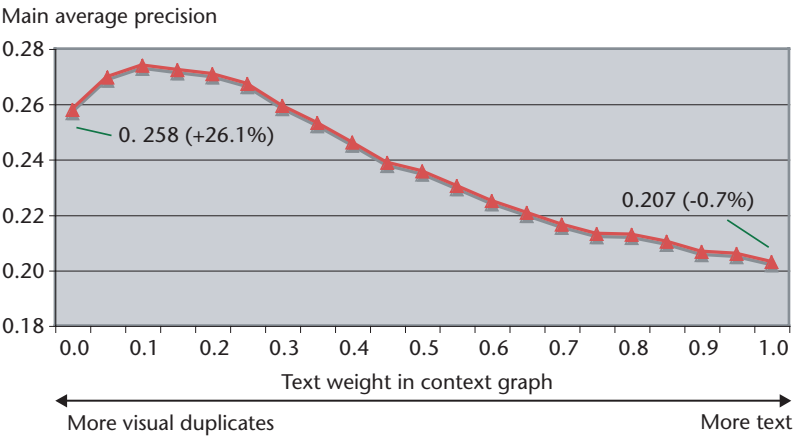

Figure 8. Context reranking on the text-okapi text search set with variant text similarity weights, where $\alpha = 0.8$ in the random walk procedure. The numbers in parentheses are the relative improvements from the text search result (in MAP). The best performance of the ratio of text to duplicates is around 0.15:0.85.

Conclusion and future work

Although it operates on the story level rather than the image level, context reranking's goal is similar to that of IB reranking in that it favors recurrent stories with high initial text retrieval scores. It doesn't fuse the feature density and (cluster) search relevance linearly, but formulates the problem as a random walk framework that generally converges and gives more weight to stories in dense clusters (in terms of multimodal similarity) having higher initial text retrieval scores.

The proposed visual reranking processes are related to approaches such as transductive learning[17] and cotraining,[18] which consider the problem of using numerous unlabeled samples to boost a learning algorithm's performance when only a small set of labeled examples is available. In the search-reranking problem, however, we can't exactly locate positive data in the initial text search results, but only have the (noisy) search relevance scores from the text modality.

In the future, we'll develop new methods to speed the reranking processes in large-scale visual search systems. Beyond the visual features used in this work, we'll also explore the use of a large set of generic concept detectors in computing shot similarity or multimedia document context. **MM**

Acknowledgments

This material is based on work funded by the US government. Any opinions, findings, and conclusions or recommendations expressed in this material are ours and don't necessarily reflect the views of the US government.

References

1. A.G. Hauptmann and M.G. Christel, "Successful Approaches in the TREC Video Retrieval Evaluations," *Proc. ACM Int'l Conf. Multimedia*, ACM Press, 2004, pp. 668-675.
2. T.-S. Chua et al., "TRECVID 2004 Search and Feature Extraction Task by NUS PRIS," *TREC Video Retrieval Evaluation Online Proc.*, 2004.

3. A. Amir et al., "IBM Research TRECVID-2005 Video Retrieval System," *TREC Video Retrieval Evaluation Online Proc.*, 2005.
4. S.-F. Chang et al., "Columbia University TRECVID-2006 Video Search and High-Level Feature Extraction," *TREC Video Retrieval Evaluation Online Proc.*, 2006.
5. W.H. Hsu, L.S. Kennedy, and S.-F. Chang, "Video Search Reranking via Information Bottleneck Principle," *Proc. ACM Int'l Conf. Multimedia*, ACM Press, 2006, pp. 35-44.
6. R. Sarukkai, "Video Search: Opportunities and Challenges," *Proc. Multimedia Information Retrieval Workshop* (MIR), ACM Press, 2005, p. 2.
7. J.G. Carbonell et al., "Translingual Information Retrieval: A Comparative Evaluation," *Proc. Int'l Joint Conf. Artificial Intelligence*, Morgan Kaufmann, 1997, pp. 708-715.
8. R. Yan, A. Hauptmann, and R. Jin, "Multimedia Search with Pseudo-Relevance Feedback," *Proc. Int'l Conf. Image and Video Retrieval*, LNCS 2728, Springer-Verlag, 2003, pp. 238-247.
9. A. Natsev, M.R. Naphade, and J. Tesic, "Learning the Semantics of Multimedia Queries and Concepts from a Small Number of Examples," *Proc. ACM Int'l Conf. Multimedia*, ACM Press, 2005, pp. 598-607.
10. W.H. Hsu and S.-F. Chang, "Topic Tracking across Broadcast News Videos with Visual Duplicates and Semantic Concepts," *Proc. Int'l Conf. Image Processing* (ICIP), IEEE Press, 2006, pp. 141-144.
11. N. Slonim and N. Tishby, "Agglomerative Information Bottleneck," *Neural Information Processing Systems Conf.* (NIPS), MIT Press, 1999, pp. 617-623.
12. W.H. Hsu, "An Information-Theoretic Framework towards Large-Scale Video Structuring, Threading, and Retrieval," PhD dissertation, Graduate School of Arts and Sciences, Columbia Univ., 2006.
13. X. Liu and W.B. Croft, "Cluster-Based Retrieval Using Language Models," *Proc. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval*, ACM Press, 2004, pp. 186-193.
14. M. Meila and J. Shi, "Learning Segmentation with Random Walk," *Neural Information Processing Systems Conf.* (NIPS), MIT Press, 2001, pp. 873-879.
15. Y. Yang et al., "Learning Approaches for Detecting and Tracking News Events," *IEEE Intelligent Systems*, vol. 14, no. 4, 1999, pp. 32-43.
16. L. Page et al., "The Pagerank Citation Ranking: Bringing Order to the Web," Stanford Digital Library Technologies Project, tech. report, 1998.
17. T. Joachims, "Transductive Inference for Text Classification Using Support Vector Machines," *Proc. 16th Int'l Conf. Machine Learning*, Morgan Kaufmann, 1999, pp. 200-209.
18. A. Blum and T. Mitchell, "Combining Labeled and Unlabeled Data with Co-Training," *Proc. Ann. Workshop Computational Learning Theory*, ACM Press, 1998, pp. 92-100.

Winston H. Hsu is an assistant professor in the Graduate Institute of Networking and Multimedia at National Taiwan University. His research interests include multimedia content analysis and image/video indexing. Hsu earned a PhD from the Department of Electrical Engineering, Columbia University. He is a member of the IEEE.

Lyndon S. Kennedy is a PhD candidate in the Electrical Engineering Department at Columbia University. His research interests include image/video indexing and searching. He is a member of the IEEE.

Shih-Fu Chang is a professor in the Electrical Engineering Department at Columbia University. His research interests include multimedia content analysis, image/video search, and multimedia forgery detection. Chang has a PhD from the University of California, Berkeley. He is a fellow of the IEEE.

Readers may contact Winston Hsu at the Graduate Institute of Networking and Multimedia, National Taiwan University, 1 Roosevelt Road, Section 4, Taipei 106, Taiwan; winston@csie.ntu.edu.tw.