CrowdReranking: Exploring Multiple Search Engines for Visual Search Reranking *

Yuan Liu [†‡], Tao Mei [‡], Xian-Sheng Hua [‡]
[†] University of Science and Technology of China, Hefei 230027, P. R. China
[‡] Microsoft Research Asia, Beijing 100190, P. R. China
yuanliu.ustc@gmail.com, {tmei,xshua}@microsoft.com

ABSTRACT

Most existing approaches to visual search reranking predominantly focus on mining information within the initial search results. However, the initial ranked list cannot provide enough cues for reranking by itself due to the typically unsatisfying visual search performance. This paper presents a new method for visual search reranking called *CrowdReranking*, which is characterized by mining relevant visual patterns from image search results of multiple search engines which are available on the Internet. Observing that different search engines might have different data sources for indexing and methods for ranking, it is reasonable to assume that there exist different search results yet certain common visual patterns relevant to a given query among those results. We first construct a set of visual words based on the local image patches collected from multiple image search engines. We then explicitly detect two kinds of visual patterns, i.e., *salient* and *concurrent* patterns, among the visual words. Theoretically, we formalize reranking as an optimization problem on the basis of the mined visual patterns and propose a close-form solution. Empirically, we conduct extensive experiments on several real-world search engines and one benchmark dataset, and show that the proposed CrowdReranking is superior to the state-of-the-art works.

Categories and Subject Descriptors

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

General Terms

Algorithms, Performance, Experimentation.

Keywords

Visual search, search reranking, data mining.

* This work was performed when Yuan Liu was visiting Microsoft Research Asia as a research intern.

SIGIR'09, July 19–23, 2009, Boston, Massachusetts, USA.

1. INTRODUCTION

The explosive growth and widespread accessibility of community contributed media contents on the Internet have led to surge of research activity in visual search [11]. Due to the great success of text search, most popular image and video search engines, such as Google [5], Yahoo! [23], and Live [16], build upon text search techniques by using the text information associated with media contents. This kind of visual search approach has proven unsatisfying as it entirely ignores the visual contents as a ranking signal.

To address this issue, search reranking has received increasing attention in recent years. It is defined as reordering visual documents based on multimodal cues to improve search performance. The documents might be images or video shots. The research on visual search reranking has proceeded along two dimensions from the perspective of the external knowledge used: *self-reranking* which only uses initial search results, and *query-example-based reranking* which leverages user-provided query examples. In this paper, we propose a new method, called *CrowdReranking*, by exploring multiple image and video search engines or sites which are available on the Internet.

The first dimension predominantly focuses on detecting the recurrent patterns solely in the initial search results, followed by using such patterns to perform reranking [6] [7] [9] [12] [20]. However, it is well-known that existing visual search engines do not have satisfying performance, mainly because of the noisy and even missing surrounding text. Therefore, the initial ranked list usually cannot provide enough cues to detect recurrent patterns for reranking. For example, it can be observed in Figure 1 that there are few relevant results in the top search results of TRECVID 2007 [15], Engine I, and Engine II [1]. It is difficult to achieve satisfying reranking if we solely mine information within the initial search results.

To address this issue, the second dimension leverages a few query examples to train the reranking models [10] [13] [17] [24]. The search performance can be improved due to the external knowledge derived from these examples. The model-based methods in this dimension assume the availability of a large collection of training samples. However, it is typical that training examples are too expensive to obtain as users are reluctant to provide enough query examples while searching.

On one hand, existing approaches have not been widely applied due to the very limited information they can mine

[1] We use Engine I, II, III, and IV to represent four popular image and video search engines to preserve anonymity.

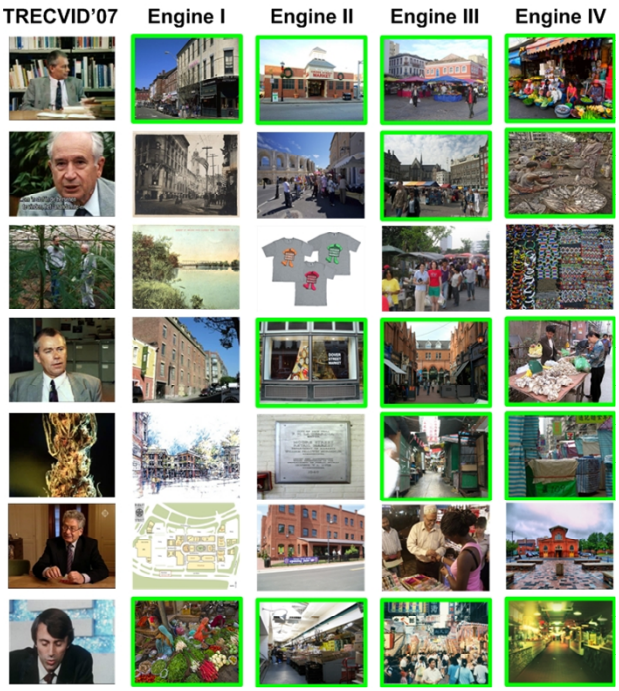


Figure 1: The exemplary top seven search results selected from TRECVID 2007 [21] and four search engines. The query is "street market scene" which is in the query list of TRECVID 2007 video search task. The green rectangles indicate relevant results. Note that the search results of TRECVID 2007 are obtained by our submission based on automatic text search [15]. [Best viewed in color]

for guiding the reranking process. On the other, there exists rich crowdsourcing knowledge available online that can be used for reranking. For example, there are a number of search engines (e.g., Google [5], Yahoo! [23], and Live [16]) and social media sites (e.g., Flickr [4]) supporting different kinds of visual search abilities. The following observations inspire the idea of leveraging search results from multiple visual search engines for reranking.

- Different search engines have different search results as they might have different data sources and metadata for indexing, as well as different search and filtering methods for ranking, as shown in Figure 1. Using search results from different engines can inform and complement the relevant visual information for each other. Thus, the reranking performance can be significantly improved due to the richer knowledge involved.
- Although a single search engine cannot always have enough cues for reranking, it is reasonable to assume that across the search results from multiple engines, there are common visual patterns relevant to a given query. For example, it would be easier to find the relevant visual patterns about "street market scene" among the results from multiple engines rather than each individual in Figure 1. The repetition in a large fraction of the images is an important signal that can be used to infer a common "visual pattern" throughout the multiple sets.

Motivated by the above observations, we propose a new method for visual search reranking, called CrowdReranking. Rather than only uses a single search engine, CrowdReranking is characterized by mining relevant visual patterns from the search results of *multiple* search engines. Given a query, finding the representative visual patterns, as well as their relative strengths and relations in multiple sets of images is the basis of CrowdReranking proposed in this study. As local features have proven effective for visual recognition in a large-scale image set [3] [9], we first construct a set of representative visual words based on the local image patches from multiple image search engines. We then explicitly detect two kinds of visual patterns relevant to the given query, i.e., *salient* and *concurrent* patterns, among the visual words. The salient pattern indicates the importance of each visual word, while the concurrent pattern expresses the interdependent relations among the visual words. The concurrent pattern, usually called context, is known to be informative for vision applications [22]. Intuitively, if a visual word is with high importance for a given query, then other words co-occurring with it would be prioritized. Therefore, we adopt a graph propagation method like PageRank [1] [9], by treating visual words as pages and their concurrence as hyperlinks. The stationary probabilities over the PageRank graph are represented as the salient pattern, while the concurrent pattern is estimated based on the propagation of the weights of graph edges. We then formalize reranking as an optimization problem which maximally preserves the initial ranked list and simultaneously matches the reranked list against the learned visual patterns as much as possible.

The remainder of this paper is organized as follows. Section 2 introduces the CrowdReranking approach. Section 3 shows experiments, followed by conclusions in Section 4.

2. CROWDRERANKING APPROACH

2.1 Overview

The objective of CrowdReranking is to mine certain visual patterns which are relevant to a given query from the search results of multiple search engines, then these patterns are used to obtain an optimal reranked document list which has the best match against the mined patterns. The flowchart of CrowdReranking is illustrated in Figure 2. Given a textual query, an initial ranked list of visual documents (images or video shots) is obtained by text search technique based on image surrounding text or video transcripts. Meanwhile, this query is fed to multiple image and video search engines or sites (e.g., Google, Yahoo!, Live, and Flickr image and video search engines) to obtain different lists of search results. First, we detect a set of representative visual words by clustering the local features of image patches which are collected from the search results of multiple engines. We then construct a graph in which the visual words are nodes and the edges between the nodes are weighted by their concurrent relations. Through a propagation process which takes the initial ranks and the reliability of search engines into account, we can explicitly detect the relevant visual patterns, including salient and concurrent patterns. The reranking is then formalized as an optimization problem on the basis of the mined visual patterns, as well as the Bag-of-Words (BoW) representation of the initial ranked list. A close-form solution can be achieved to this optimization problem.

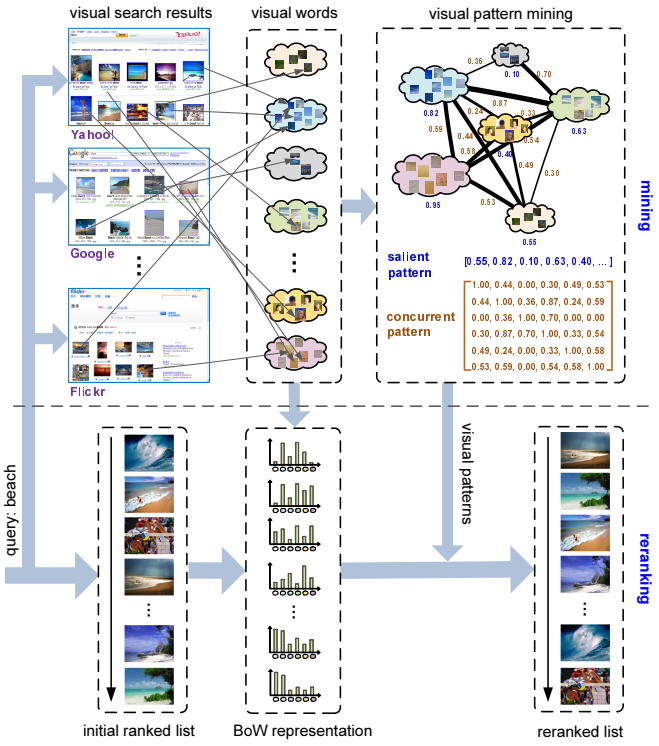


Figure 2: The flowchart of CrowdReranking.

2.2 Problem Formulation

Suppose we have a document set $\mathcal{X}$ with N documents to be reranked, where $\mathcal{X} = \{x_1, x_2, \ldots, x_N\}$. Let $\bar{r}_i$ and r_i denote the initial ranking score (i.e., relevance) and reranking score for document x_i. In the reranking problem, the initial ranking scores are to be preserved as they indicate the relevance information from text perspective. On the other hand, the reranked list should be consistent with the learned knowledge (i.e., visual patterns) from multiple search engines. Therefore, we can formulate the reranking problem by minimizing the following energy function:

$$E(\mathbf{r}) = Dist(\mathbf{r}, \bar{\mathbf{r}}) - \lambda Cons(\mathbf{r}, \mathcal{K}) \quad (1)$$

where $\bar{\mathbf{r}} = [\bar{r}_1, \bar{r}_2, \ldots, \bar{r}_N]^{\mathrm{T}}$ and $\mathbf{r} = [r_1, r_2, \ldots, r_N]^{\mathrm{T}}$. $Dist(\mathbf{r}, \bar{\mathbf{r}})$ corresponds to the *ranking distance*, while $Cons(\mathbf{r}, \mathcal{K})$ corresponds to the *consistence* between the reranked list $\mathbf{r}$ and the learned knowledge $\mathcal{K}$. $\mathcal{K}$ indicates the learned knowledge from multiple search engines which also corresponds to the mined visual patterns in this paper. The parameter λ tunes the contribution of knowledge $\mathcal{K}$ to the reranked list. When $\lambda = 0$, the reranked list $\mathbf{r}$ will be the same as the initial ranked list.

2.3 Visual Pattern Mining

As we look for common visual patterns across different ranked lists of images, the pattern representations should be invariant to a variety of degradations (scale, orientation, global or local appearance, and so on). We adopt scale-invariant feature transform (SIFT) descriptor with a Difference of Gaussian (DoG) interest point detector in this work, as it has proven to be effective for large-scale visual recognition [9] [14]. The interest point is referred to as local salient patch in this paper, each associated with a 128-dimensional feature vector. We further adopt K-Means to cluster the similar patches into "visual words," and use Bag-of-Words (BoW) to represent each image [19]. For a given query, the visual patterns $\mathcal{K}$ will be mined from the visual words collected from the search results of multiple search engines.

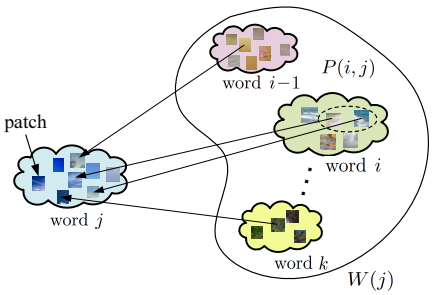


Figure 3: The computation of visual patterns.

Specifically, we investigate two kinds of visual patterns in this work: *salient* and *concurrent* patterns. The salient pattern indicates the importance of each visual word, while concurrent pattern expresses the interdependent relations among the visual words. The premise of using concurrence as hyperlinks is that if a visual word is viewed important, then other co-occurring or similar visual words also might be of interest. For example, for a query "beach," visual words extracted at the "sea" patch is ranked high, then the co-occurring "sand" and "sky" patches should be also prioritized. Therefore, we adopt PageRank-like propagation framework and construct a graph with the visual words as nodes and co-occurrence between the visual words as hyperlinks. Suppose we have L visual words, the visual pattern $\mathcal{K}$ is expressed as the combination of salient pattern $\mathbf{q}$ and concurrent pattern $\mathbf{C}$:

$$\mathcal{K} \triangleq \mathcal{K}(\mathbf{q}, \mathbf{C}) \quad (2)$$

where $\mathbf{q} = [q_1, q_2, \ldots, q_L]^{\mathrm{T}}$ is a L-dimensional vector with each element indicating the salience or importance of a visual word, and $\mathbf{C} = [c_{mn}]_{(L \times L)}$ is a $L \times L$ matrix with each element indicating the hyperlink between two words.

Let $W(j)$ denote the set of words that contain patches connecting to the patches in word j, and $P(i,j)$ denote the set of patches in word i connecting to the patches in word j, as shown in Figure 3. The salience of word j after the k-th iteration, $q_j(k)$, is given in a way similar to PageRank [1]:

$$q_j(k) = \varepsilon q_j(0) + (1-\varepsilon) \sum_{i \in W(j)} \frac{|P(i,j)|}{\sum_{k=1}^{L} |P(i,k)|} q_i(k-1) \quad (3)$$

where $|\cdot|$ denotes the size of a set, ε $(0 < \varepsilon < 1)$ is the weight balancing the initial and the propagated salience scores. $q_j(0) = \sum_\ell x_j^\ell$, x_j^ℓ denotes the normalized ranking score [7] of the ℓ-th patch from word j in the initial ranked list.

Accordingly, the concurrent pattern is given by the average weight between word i and j over the graph:

$$c_{ij} = \left(\frac{|P(i,j)|}{\sum_{k=1}^{L} |P(i,k)|} + \frac{|P(j,i)|}{\sum_{k'=1}^{L} |P(j,k')|} \right) \times 0.5 \quad (4)$$

2.4 Reranking

This section discusses the ranking distance $Dist(\mathbf{r}, \bar{\mathbf{r}})$ and consistence $Cons(\mathbf{r}, \mathcal{K})$ based on the mined knowledge $\mathcal{K}(\mathbf{q}, \mathbf{C})$. Recently, some researchers have proposed various ranking distances, mainly including pointwise and pairwise distances as follows.

- Pointwise ranking distance [6]:

$$Dist(\mathbf{r}, \bar{\mathbf{r}}) = \sum_n (r_n - \bar{r}_n)^2 \tag{5}$$

- Pairwise ranking distance [20]:

$$Dist(\mathbf{r}, \bar{\mathbf{r}}) = \sum_{m,n} \left(1 - \frac{r_m - r_n}{\bar{r}_m - \bar{r}_n}\right)^2 \tag{6}$$

For the consistency, most existing reranking methods solely use the visual consistency within the initial search results, assuming that visual documents similar in appearance are with similar ranks [6] [20]. As we have mentioned, only mining within initial ranked list is not reasonable when there exist much more irrelevant documents than relevant ones. In this work, we leverage the mined visual pattern $\mathcal{K}$ to define a more suitable consistence.

Let $\mathbf{f}_n = [f_{n1}, f_{n2}, \ldots, f_{nL}]^{\mathrm{T}}$ denote the BoW representation for image x_n [19], the consistence is defined by

$$Cons\,(\mathbf{r}, \mathcal{K}) = \sum_n \left(\sum_i q_i f_{ni} + \sum_{i,j} c_{ij} f_{ni} f_{nj}\right) r_n \tag{7}$$

Let $\mathbf{s} = [s_1, s_2, \ldots, s_N]^{\mathrm{T}}$ denote the vector with entries $s_n = \sum_i q_i f_{ni} + \sum_{i,j} c_{ij} f_{ni} f_{nj}$. $\mathbf{s}$ can be viewed as the cosine similarity between the visual representation of image x_n and the mined visual patterns. Based on the two types of ranking distances, we integrate the above two ranking distances in equation (5) and (6), as well as the consistence in (7), to equation (1), and have the following two objective reranking functions.

- Reranking function using pointwise ranking distance:

$$\min_{\mathbf{r}} \left\{ \sum_n (r_n - \bar{r}_n)^2 - \lambda \sum_n \Big(\sum_i q_i f_{ni} + \sum_{i,j} c_{ij} f_{ni} f_{nj}\Big) r_n \right\} \tag{8}$$

We called this optimization problem as *pointwise mining-based reranking*. We can obtain the solution of Equation (8) as follows (proven in Appendix):

$$\mathbf{r} = \frac{1}{2}(2\bar{\mathbf{r}} + \lambda \mathbf{s}) \tag{9}$$

Obviously, equation (9) consists of two parts, i.e., $\bar{\mathbf{r}}$ and $\mathbf{s}$, which corresponds to the initial ranked list and the learned knowledge, respectively. Therefore, the pointwise reranking can be also viewed as the linear fusion between the initial ranked list and the ranked list learned from the online sources.

- Reranking function using pairwise ranking distance:

$$\min_{\mathbf{r}} \left\{ \sum_{m,n} \Big(1 - \frac{r_m - r_n}{\bar{r}_m - \bar{r}_n}\Big)^2 - \lambda \sum_n \Big(\sum_i q_i f_{ni} + \sum_{i,j} c_{ij} f_{ni} f_{nj}\Big) r_n \right\} \tag{10}$$

Algorithm 1 The CrowdReranking algorithm.

Input: The initial ranked list $\bar{\mathbf{r}} = [\bar{r}_1, \bar{r}_2, ..., \bar{r}_N]^{\mathrm{T}}$.
Output: The reranked list $\mathbf{r} = [r_1, r_2, ..., r_N]^{\mathrm{T}}$.
Algorithm:
1: Collect data from multiple search engines by the same query.
2: Extract SIFT features for each image and construct the visual words by K-Means.
3: Calculate salient and concurrent patterns by equation (3) and (4).
4: Obtain the reranked list according to equation (9) or (11).

We called this optimization problem as *pairwise mining-based reranking*. The solution of equation (10) with a constraint $r_N = 0$ can be derived as (proven in Appendix):

$$\mathbf{r} = \frac{1}{2}\breve{\mathbf{\Delta}}^{-1}(2\breve{\mathbf{c}} + \lambda \breve{\mathbf{s}}) \tag{11}$$

where $\mathbf{c} = \mathbf{2}(\mathbf{U}\mathbf{e})^{\mathrm{T}}$, $\breve{\mathbf{c}}$ and $\breve{\mathbf{s}}$ are obtained by replacing the last element of $\mathbf{c}$ and $\mathbf{s}$ with zero, respectively. $\mathbf{\Delta} = \mathbf{D} - \mathbf{U}$, where $\mathbf{U} = [u_{mn}]_{(N \times N)}$ denotes an anti-symmetric matrix with $u_{mn} = \frac{1}{\bar{r}_m - \bar{r}_n}$, and $\mathbf{D}$ is a diagonal matrix with its $(n\text{-}n)$-element $d_{nn} = \sum_{n=1}^{N} u_{mn}$. $\breve{\mathbf{\Delta}}$ is obtained by replacing the last row of $\mathbf{\Delta}$ with $[0, \ldots, 0, 1]_{(1 \times N)}$.

In equation (11), there are also two parts, i.e., $\breve{\mathbf{\Delta}}^{-1}\breve{\mathbf{c}}$ and $\breve{\mathbf{\Delta}}^{-1}\breve{\mathbf{s}}$, in which $\breve{\mathbf{\Delta}}^{-1}\breve{\mathbf{c}}$ is solely determined by the initial ranked list, while $\breve{\mathbf{\Delta}}^{-1}\breve{\mathbf{s}}$ can be viewed as the learned knowledge biased by the initial ranked list. Therefore, the reranked list can be also viewed as the combination of the initial ranked list and the learned knowledge.

In summary, we get the flowchart of CrowdReranking in Algorithm 1.

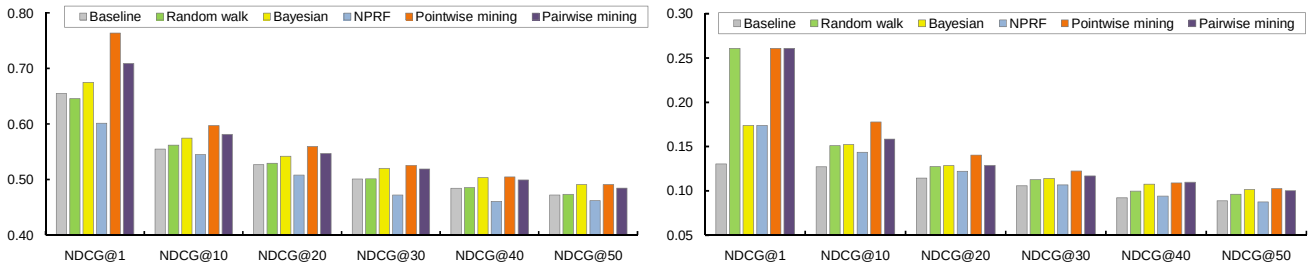
3. EXPERIMENTS

3.1 Data

We conducted experiments on two image datasets. One is real-world image data collected from three popular image search engines (i.e., Engine I, II, and III) and a photo sharing site (i.e., Engine IV, also called "Web set" in short). We selected 29 representative top queries from the query history of one popular search engine. These queries consist of a variety of types, including objects, people, event, entertainments, location, and time [2]. For each query, we collected about top 1,000 images returned by each search engine. After some parts of unaccessible searched images are filtered, the dataset contains 74,000 images in total, and the ranks of these images are kept as the initial ranked list.

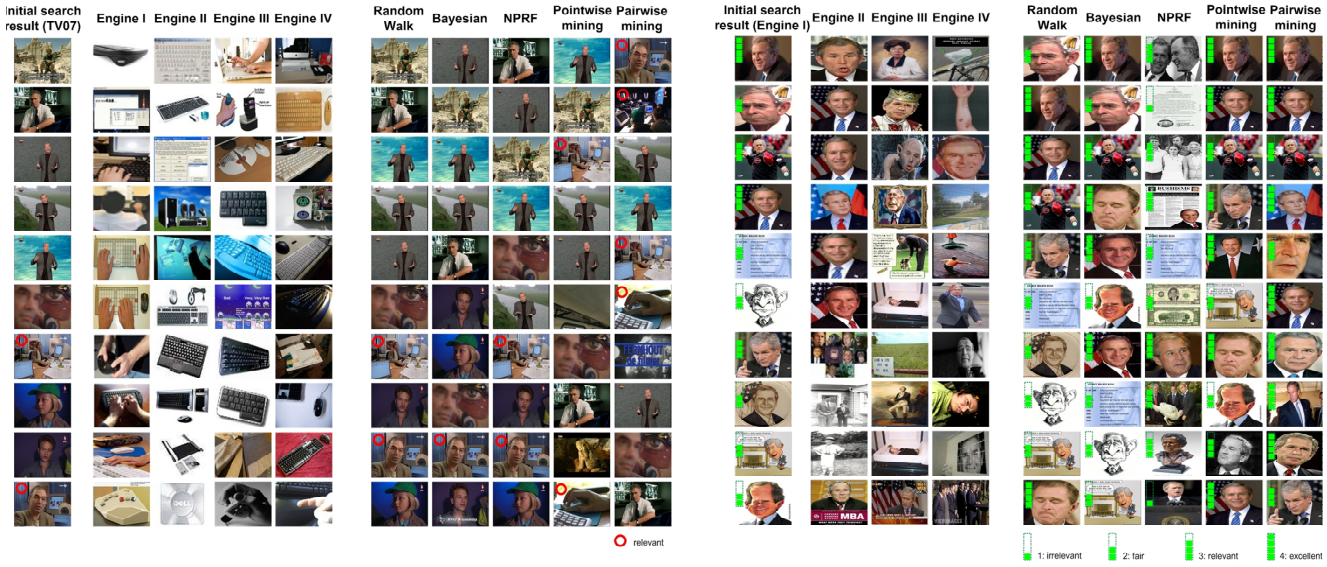
The other dataset is the benchmark TRECVID 2007 test set [21] (called "TV07 set" in short), which consists of 18,142 video shots. The search and reranking are performed on the basis of the keyframe and transcript for each shot. There are 24 text queries with their ground truth of relevance. The description of each query can be found in [21]. The text

[2] The queries include: (1) animal, (2) beach, (3) Beijing Olympic 2008, (4) building, (5) car, (6) cat, (7) clouds, (8) earth, (9) flower, (10) fox, (11) funny dog, (12) George W. Bush, (13) grape, (14) hearts, (15) hello kitty, (16) hiking, (17) Mercedes logo, (18) panda, (19) sky, (20) statue of liberty, (21) sun, (22) trees, (23) wedding, (24) white cat, (25) white house, (26) white house night, (27) winter, (28) yellow rose, and (29) zebra.



(a) Web set (b) TV07 set

Figure 4: Comparison of reranking methods in terms of NDCG. Results in Web set are the average of three search engines.



(a) (b)

Figure 5: Examples of different methods for two queries. (a) TRECVID 2007 query: "Find shots of hands at a keyboard typing or using a mouse." (b) Web query: "George W. Bush." [Best viewed in color]

search results by Okapi BM25 [18] were used as the initial ranked list in the following experiments.

We can find that most queries in the real dataset are represented as a general word or a simple phrase, while those in TV07 usually describe an event or a scene with much more words. The TV07 has much lower text search performance as the queries are more challenging. Therefore, the two datasets can be viewed as representative and complementary datasets in both research and real applications.

3.2 Methodologies

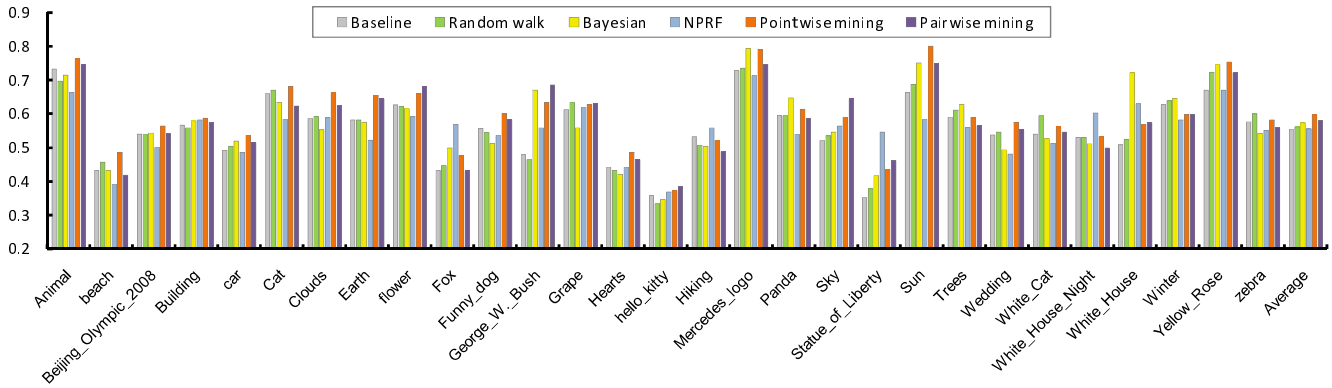
For the Web set from the four sites, the relevance of each returned image to the corresponding query was manually labeled by three subjects on a 1-4 scales: (1) "irrelevant," (2) "fair," (3) "relevant," and (4) "excellent." The ground truth relevance of each image is the median scale of the three evaluations. We adopt NDCG as performance metric since it is widely used to deal with multiple relevance levels [8]. Given a query q, the NDCG score at the depth d in the ranked documents is defined by

$$NDCG@d = Z_d \sum_{j=1}^{d} \frac{2^{r^j} - 1}{\log(1+j)} \quad (12)$$

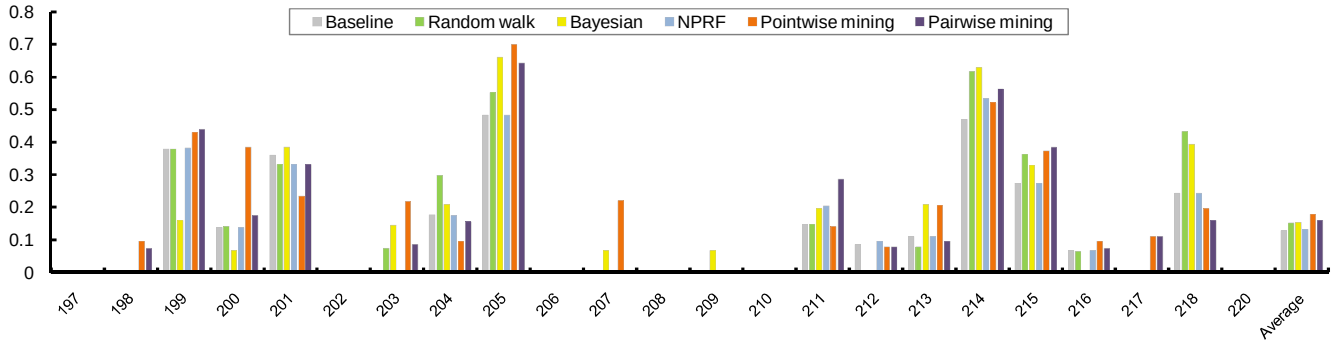
where r^j is the rating of the j-th document, Z_d is a normalization constant and is chosen so that a perfect ranking's $NDCG@d$ value is 1. For TV07 set, we also used the NDCG to evaluate the performance based on the available relevance (i.e., two scales of "positive" or "negative").

In our experiments, for the Web set, we reranked the search result of each search engine by using the other three engines. For TV07 set, we used the mined visual patterns from all four search engines. We used top 100 images from each engine for visual pattern mining and top 500 images in the initial search results for reranking, since it is typical that there are very few relevant images after the top 500 search results. The number of visual words is empirically set to 2,000 [19].

To demonstrate the effectiveness of the proposed CrowdReranking methods which include *pointwise mining-based reranking* and *pairwise mining-based reranking*, we compared with the following three state-of-the-art reranking methods.



(a) Web set



(b) TV07 set

Figure 6: Performance of each query in Web set and TV07 set measured by $NDCG@10$. Results in Web set are the average of three search engines.

For all these approaches, we select parameters according to their globally best performance setting in our experiments.

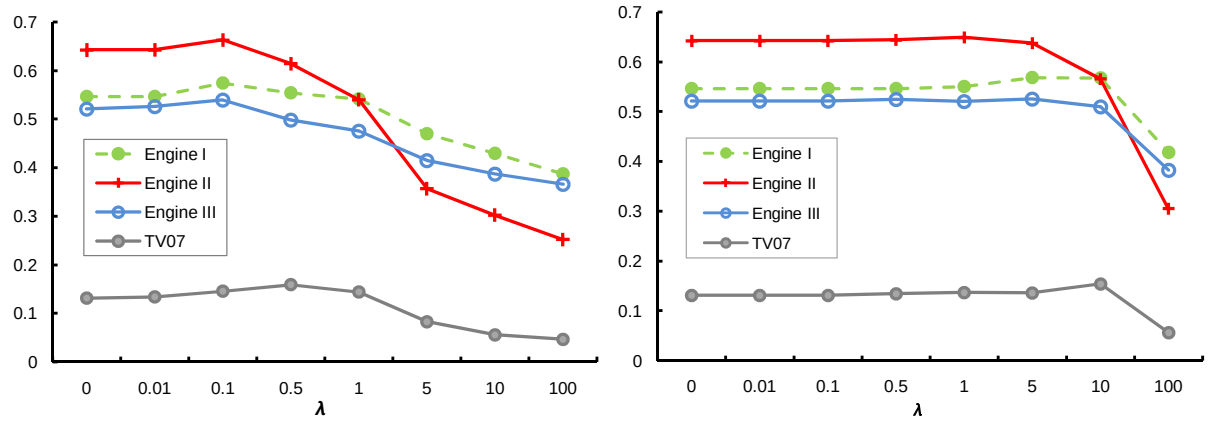
- Random walk reranking [6].
- Bayesian reranking [20]. The first two are representative methods in the first research dimension for visual search reranking (i.e., self-ranking).
- NPRF reranking [24]. The third is a representative method in the second research dimension (i.e., query-example-based reranking). It uses the query examples as "positive" documents and randomly samples the low-rank images as "negative" to train the SVM models based on global image features. To fairly compare with our approach, we directly use the top-ranked images from the multiple search engines as the query examples.

3.3 Evaluations

3.3.1 *Evaluation of Reranking Performance*

The experimental results are shown in Figure 4 and 5, from which we can see that the proposed two reranking approaches outperform the others. Moreover, it can be observed that:

- The improvements of the proposed CrowdReranking over Random walk and Bayesian reranking methods indicate that mining external knowledge, especially the visual patterns from the search results of multiple search engines, can benefit reranking a lot.
- The superiority of the proposed CrowdReranking to NPRF-based reranking indicates that the mined visual patterns from the local visual words are more effective than the global features discovered solely from top-ranked images for reranking.
- NPRF reranking did not outperform Random walk reranking, Bayesian reranking, as well as CrowdReranking. The reasons are two-fold: (1) Rather than the object-related queries in the Web set, the queries in TV07 set are more specific to event and scene, which in general represent more diverse appearance. Therefore, it is difficult to discover the common visual patterns from external sources. In contrast, the proposed CrowdReranking which mines the patterns from visual words on the basis of local features can effectively leverage the external sources and thus achieve better performance. (2) The "negative" samples randomly selected from the low-ranked images have diverse appearances, which makes the models in NPRF-based reranking do not have a good generalization ability.
- Among these reranking methods, Random walk and Bayesian reranking respectively has three parameters (i.e., the number of nearest-neighbors, the trade-off parameter and the scaling parameter in Gaussian kernel), and about 400 sets of parameters in total to select the



(a) Pointwise mining-based

(b) Pairwise mining-based

Figure 7: Performance with different λ measured by NDCG@10. Note that "$\lambda = 0$" corresponds to the initial search without reranking.

optimal one; while NPRF has two parameters (i.e., C and γ in RBF kernel of SVM), and 70 sets of parameters in total. In addition, different sets of parameters are elaborately selected for different search engines. As a result, the robustness of these methods is limited. In contrast, the proposed CrowdReranking has only one parameter λ. From the Figure 7, we can see that it reaches the peak at a relatively stable value for different search engines.

Furthermore, the performance improvements are consistent and stable—most queries are improved compared to the initial ranked lists and have better performance than the other methods, as shown in Figure 6. The performance of some queries has significant improvement even their initial search results are extremely poor, such as the query of "(198) a door being opened" and "(217) a road taken from a moving vehicle through the front windshield." However, we can see that the performance of some queries degrades with the random walk and Bayesian reranking. This indicates that only mining within initial ranked list could not be always enough to obtain satisfying reranking results. On the other hand, NPRF-based reranking, which also mines the external online sources, is not stable in terms of reranking.

Figure 5 shows the top 10 images of different engines and reranking approaches. It is difficult to discover relevant visual patterns solely from the initial search results for the given query in TV07 set as there are only two samples somewhat relevant. However, based on the search results from multiple search engines, we can mine salient and concurrent visual patterns about the query. As a result, the relevant documents can be ranked higher in the reranked list.

3.3.2 *Evaluation of Trade-off Parameter λ*

We also investigated the performance of CrowdReranking with different tradeoff parameter λ in equation (1). Figure 7 shows the performance of the pointwise and pairwise mining-based reranking methods with different λ in terms of $NDCG@10$. From the figures, we can see that the performance curve is like a "Λ" shape as λ increases.

- As shown in Figure 7 (a), the performance of pointwise mining-based reranking increases when λ increases and arrives at the peak at $\lambda = 0.1$ on Web set, while it reaches the peak at $\lambda = 0.5$ on TV07 set. As aforementioned, TV07 set has more challenging queries (i.e., long and complex queries) and thus has worse performance compared with Web set. Basically, a relatively larger λ would be more suitable for a worse initial search results as in TV07 set. It can be concluded that the trade-off parameter λ can be set according to the performance of initial search results. The same conclusion can be drwan in the pairwise mining-based reranking, as shown in Figure 7 (b).
- When λ goes to infinity, the reranking process relies almost entirely on the learned knowledge with the initial search results excluded. The reranking will reduce to the query-by-example (QBE) search problem and the performance will significantly degrade. From this observation, we can conclude that both the initial search results and external knowledge play important roles in the reranking. However, it remains an open problem on how to find an optimal λ.

4. CONCLUSIONS

In this paper, we have proposed a novel visual reranking method by mining relevant visual patterns from the search results which are available from existing search engines on the Internet. To the best of our knowledge, the proposed CrowdReranking represents the first attempt towards leveraging crowdsourcing knowledge for visual reranking.

There are several open problems for further studies. First, the number of search engines or online resources is still limited in this work. It would be a promising topic to discover more search engines and sites and investigate how many engines and sites are enough for reranking. Second, most of current search engines are multilingual systems. We can use the results from multilingual systems for reranking. Third, the well-organized online knowledge like Wikipedia and Mediapedia can be also leveraged for visual reranking.

5. REFERENCES

[1] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, 30(1-7):107–117, 1998.

[2] L. B. Cremeant and R. M. Murra. Stability analysis of interconnected nonlinear systems under matrix feedback. In *Proceedings of IEEE Conference on Decision and Control*, 2003.
[3] R. Fergus, L. Fei-Fei, P. Perona, and A. Zisserman. Learning object categories from Google's image search. In *Proceedings of IEEE International Conference on Computer Vision*, 2005.
[4] Flickr. http://www.flickr.com/.
[5] Google image and video search. http://image.google.com/, http://video.google.com/.
[6] W. Hsu, L. Kennedy, and S.-F. Chang. Video search reranking through random walk over document-level context graph. In *Proceedings of ACM International Conference on Multimedia*, Augsburg, Germany, 2007.
[7] W. H. Hsu, L. S. Kennedy, and S.-F. Chang. Video search reranking via information bottleneck principle. In *Proceedings of the ACM International Conference on Multimedia*, Santa Barbara, USA, 2006.
[8] K. Jarvelin and J. Kekalainen. IR evaluation methods for retrieving highly relevant documents. In *Proceedings of ACM SIGIR*, 2000.
[9] Y. Jing and S. Baluja. PageRank for product image search. In *Proceedings of International World Wide Web Conference*, 2008.
[10] L. Kennedy and S.-F. Chang. A reranking approach for context-based concept fusion in video indexing and retrieval. In *Proceedings of ACM International Conference on Image and Video Retrieval*, 2007.
[11] M. S. Lew, N. Sebe, C. Djeraba, and R. Jain. Content-based multimedia information retrieval: State of the art and challenges. *ACM Transactions on Multimedia Computing, Communications and Applications*, 2(1):1–19, February 2006.
[12] Y. Liu, T. Mei, X.-S. Hua, J. Tang, X. Wu, and S. Li. Learning to video search rerank via pseudo preference feedback. In *Proceedings of IEEE International Conference on Multimedia & Expo*, 2008.
[13] Y. Liu, T. Mei, X. Wu, and X.-S. Hua. Optimizing video search reranking via minimum incremental information loss. In *Proceedings of ACM International Workshop on Multimedia Information Retrieval*, 2008.
[14] D. Lowe. Object recognition with informative features and linear classification. In *Proceedings of IEEE International Conference on Computer Vision*, 2003.
[15] T. Mei, X.-S. Hua, W. Lai, L. Yang, and et al. MSRA-USTC-SJTU at TRECVID 2007: High-level feature extraction and search. In *TREC Video Retrieval Evaluation Online Proceedings*, 2007.
[16] Microsoft Live image and video search. http://www.live.com/?scope=images/, http://www.live.com/?scope=videos/.
[17] A. Natsev, A. Haubold, J. Tešić, L. Xie, and R.Yan. Semantic concept-based query expansion and re-ranking for multimedia retrieval. In *Proceedings of ACM International Conference on Multimedia*, 2007.
[18] S.-E. Robertson and K.-S. Jones. Simple, proven approaches to text retrieval. *Cambridge University Computer Laboratory Technical Report TR356*, 1997.
[19] J. Sivic and A. Zisserman. Video Google: A text retrieval approach to object matching in videos. In *Proceedings of IEEE International Conference on Computer Vision*, 2003.
[20] X. Tian, L. Yang, J. Wang, Y. Yang, X. Wu, and X.-S. Hua. Bayesian video search reranking. In *Proceedings of ACM International Conference on Multimedia*, 2008.
[21] TRECVID. http://www-nlpir.nist.gov/projects/trecvid/.
[22] L. Wolf and S. Bileschi. A critical view of context. *International Journal of Computer Vision*, 9(2):251–261, 2006.
[23] Yahoo image and video search. http://image.yahoo.com/, http://video.yahoo.com/.
[24] R. Yan, A. Hauptmann, and R. Jin. Multimedia search with pseudo-relevance feedback. In *Proceedings of ACM International Conference on Image and Video Retrieval*, 2003.

6. APPENDIX

6.1 Pointwise mining-based reranking

Rewrite equation (8) in the matrix way:

$$\min_{\mathbf{r}} \left\{ (\mathbf{r}-\bar{\mathbf{r}})^{\mathrm{T}}(\mathbf{r}-\bar{\mathbf{r}}) - \lambda \mathbf{s}^{\mathrm{T}}\mathbf{r} \right\} \tag{13}$$

Then, taking derivatives and equate it to zero, we can obtain

$$2(\mathbf{r}-\bar{\mathbf{r}}) = \lambda \mathbf{s} \tag{14}$$

Then, the solution is

$$\mathbf{r} = \frac{1}{2}(2\bar{\mathbf{r}} + \lambda \mathbf{s}). \tag{15}$$

6.2 Pairwise mining-based reranking

Revisit equation (10) as follows:

$$\begin{aligned} &\min_{\mathbf{r}} \sum_{m,n} (1 - \tfrac{r_m - r_n}{\bar{r}_m - \bar{r}_n})^2 - \lambda \sum_n (\sum_i q_i f_{ni} + \sum_{i,j} c_{ij} f_{ni} f_{nj}) r_n \\ &= \min_{\mathbf{r}} \sum_{m,n} u_{mn}^2 (r_m - r_n)^2 - 2 \sum_{m,n} u_{mn}(r_m - r_n) - \\ &\lambda \sum_n s_n r_n + const \end{aligned}$$

Rewrite it in the matrix way:

$$\min_{\mathbf{r}} \left\{ 2\mathbf{r}^{\mathrm{T}}\mathbf{\Delta}\mathbf{r} - \mathbf{2c}^{\mathrm{T}}\mathbf{r} - \lambda \mathbf{s}^{\mathrm{T}}\mathbf{r} \right\} \tag{16}$$

where $\mathbf{\Delta} = \mathbf{D} - \mathbf{U}$, $\mathbf{U} = [u_{mn}]_{(N \times N)}$ denotes an anti-symmetric matrix with $u_{mn} = \frac{1}{\bar{r}_m - \bar{r}_n}$, $\mathbf{D}$ is a diagonal matrix with its $(n\text{-}n)$-element $d_{nn} = \sum_{m=1}^{N} u_{nm}$, and $\mathbf{c} = 2(\mathbf{U}\mathbf{e})^{\mathrm{T}}$. Taking derivatives and equate it to zero, we can obtain

$$2\mathbf{\Delta}\mathbf{r} = 2\mathbf{c} + \lambda \mathbf{s} \tag{17}$$

The solution of equation (17) is non-unique since the Laplacian matrix $\mathbf{\Delta}$ is singular [2]. Referring to [20], we also simply add a constraint $r_N = 0$ where N is the length of $\mathbf{r}$. Thus we replace the last row of $\mathbf{\Delta}$ with $[0, 0, \ldots, 0, 1]_{1 \times N}$ to obtain $\breve{\mathbf{\Delta}}$, the last element of $\mathbf{c}$ and $\mathbf{s}$ with zero to obtain $\breve{\mathbf{c}}$ and $\breve{\mathbf{s}}$, respectively. Then, the solution is

$$\mathbf{r} = \frac{1}{2}\breve{\mathbf{\Delta}}^{-1}(2\breve{\mathbf{c}} + \lambda \breve{\mathbf{s}}). \tag{18}$$