Image search results refinement via outlier detection using deep contexts*

Junyang Lu† Jiazhen Zhou† Jingdong Wang‡ Tao Mei‡ Xian-Sheng Hua§ Shipeng Li‡
†Peking University ‡Microsoft Research Asia §Microsoft Bing
{lujunyang, jiazhenzhou}@pku.edu.cn {jingdw, tmei, xshua, spli}@microsoft.com

Abstract

Visual reranking has become a widely-accepted method to improve traditional text-based image search results. The main principle is to exploit the visual aggregation property of relevant images among top results so as to boost ranking scores of relevant images, by explicitly or implicitly detecting the confident relevant images, and propagating ranking scores among visually similar images. However, such a visual aggregation property does not always hold, and thus these schemes may fail.

In this paper, we instead propose to filter out the most probable irrelevant images using deep contexts*, which is the extra information that is not limited in the current search results. The deep contexts for each image consist of sets of images that are returned by searches using the queries formed by the textual context of this image. We compare the popularity of this image in the current search results and the deep contexts to check the irrelevance score. Then the irrelevance scores are propagated to the images whose useful textual context is missed. We formulate the two schemes together to reach a Markov random field, which is effectively solved by graph cuts. The key is that our scheme does not rely on the assumption that relevant images are visually aggregated among top results and is based on the observation that an outlier under the current query is likely to be more popular under some other query. After that, we perform graph reranking over filtered results to reorder them. Experimental results on the INRIA dataset show that our proposed method achieves significant improvements over previous approaches.*

1. Introduction

Image search and retrieval has been one of the active and challenging areas of multimedia research. Most image search engines provide results based mainly on matching textual information associated with images and the input query. However, a large proportion of the search results may not be related to the query because the textual information may fail to capture image contents.

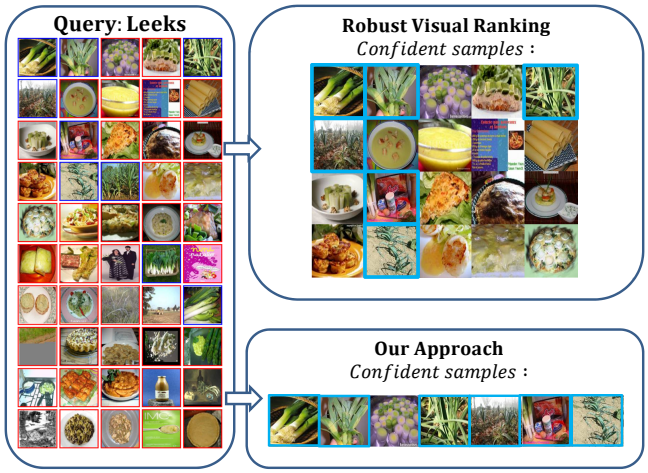

Figure 1. When irrelevant images hold dominant positions in text-based search results, previous methods detect relevant samples unsatisfactorily. Our approach makes use of deep contexts to detect outliers and generates better confident relevant samples.

In order to improve search results, a lot of efforts were conducted in designing techniques that use both textual information and visual contents to refine the search results [8, 9, 15, 20, 21, 22]. The basic principle is that visually similar images tend to be ranked together. Early methods take top-ranked results as pseudo-positive samples and bottom-ranked results as pseudo-negative samples. These samples are used to train a classifier [15, 21, 22] to compute the ranking scores or used as sources whose ranking scores are propagated to other images [8, 9, 20]. These methods are not guaranteed to work well because there always exist false positive images in the pseudo-positive samples.

Recent works on search results refinement instead investigate techniques to select relevant samples with high confidence scores and subsequently match them with the rest images to reorder search results [14, 18]. The criterion of selecting confident samples is checking if the images are visually dominant in the search results. Robust visual reranking [18] selects confident samples using the additional spar-

*This work was done when Junyang Lu and Jiazhen Zhou were interns at Microsoft Research Asia.

sity and ranking constraint, and then performs a kernel-based reranking. Noise resistant graph ranking [14] uses a spectral filter to select confident samples, and then applies a graph-based ranking algorithm. However, the information exploited by all previous methods is limited in the search results under the current query. As a result, when irrelevant samples are visually aggregated among top results rather than relevant ones, these techniques will generate false confident samples and lead to worse results.

Figure 1 shows an example in which the images visually well aggregated among top results are actually irrelevant. Query "leeks" should refer to a vegetable. However, in the text-based image search results, there are many irrelevant images related to cuisines with leeks. Since images related to cuisines appear in high initial ranks quite frequently, existing confident sample detection methods, including Robust visual reranking [18], are unable to eliminate them from confident samples.

Figure 2 shows another example where irrelevant images are visually uneasily differentiated from relevant images and also have high initial ranks. In common sense, "Arc De Triomphe" ("Triumphal Arch" in English) usually refers to the Triumphal Arch in France. However, the 4th image in "Arc De Triomphe" is actually "Volubilis Arc De Triomphe", which is located in Morocco. Though human can distinguish the 4th image from the others easily, it is not so easy for traditional reranking algorithms, since the 4th image not only looks like other relevant images, but also ranks highly in the initial ranking list. Previous state-of-art methods are unable to detect them as irrelevant images.

After a further investigation on the search results, we found that such two cases are not just exceptions. Instead, they are prevalent problems existing in unsatisfactory search results. Therefore, we realize that visual reranking with information limited in the current query is far from enough. In order to build a more reliable confident sample set, we explore other information derived from current search results.

In this paper, we propose to make use of *deep contexts* to detect irrelevant images. To collect deep contexts, we extract the textual context associated with each image in the search results for the current query. The textual context of an image is used to form a set of text queries, with which as queries we get sets of search results. Each set of search results, except the current search results, form a deep contexts of this image, associated with the corresponding query. To check if this image is irrelevant, we compare the popularity of this image in the current search results and the deep contexts and get the irrelevance (outlier) score. Furthermore, it is observed that the textual context of an image is sometimes incomplete and may miss useful words. To deal with this issue, we propose to propagate the outlier scores to visually similar images whose useful textual context is missed. Next, we formulate the two schemes together as a submodular pair-wise Markov random field (MRF) problem that can be efficiently solved by graph cuts, to identify irrelevant images. Figure 1 gives an instance of confident samples generated by eliminating outliers using our approach.

Then, we present a graph reranking approach based on the detection results to refine the search results. The graph reranking scheme makes our approach work satisfactorily for the cases that relevant images are visually aggregated, which the workability of the previous approaches often depends on. Our outlier detection scheme makes our approach also work well in the case that relevant images among top results are not visually aggregated while previous approaches usually fail.

The main contributions are summarized as follows,

- We propose to make use of *deep contexts* to detect irrelevant images, which does not rely on the assumption that relevant images among top results are visually aggregated.
- We further detect more irrelevant images by propagating the outliers detected from deep contexts to images whose useful textual context is missed.
- We formulate the irrelevant image identification problem as an MRF problem that is submodular and thus can be effectively solved by graph cuts.

2. Related works

Existing methods on image or video search reranking can be roughly divided into several categories: supervised methods [11, 23], classification-based methods [15, 21, 22], clustering-based methods [1, 2, 16], topic models [3, 4, 6], and graph-based methods [8, 9, 20].

Supervised methods learn reranking models offline from training search results that have manually annotated ranking levels and apply it on online search results [7, 11, 23]. [23] learns a query-independent reranking model with query-dependent reranking features. [11] trains an SVM classifier with both textual and visual features to get a query relative model. [7] introduces a passive-agressive model for image retrieval (PAMIR), which adopts a learning criterion related to the final retrieval performance. Experiments in [14] showed that these methods have severe over-fitting problems.

Classification-based methods first take top-ranked search results as pseudo-positive samples and bottom-ranked results as pseudo-negative samples. Then a classifier such as SVM [22], boosting [21] or ranking SVM [15], is trained from these samples to generate the ranking scores. However, pseudo-positive samples may be false positive samples and thus lead to bad results.

Clustering-based methods use visual similarity to seek clusters in search results [1, 2]. Then the images are reranked based on their distances to the cluster centers. Some works also take original ranking into consideration [16]. The drawbacks of this kind of methods include that irrelevant images may also form clusters, especially when their original ranking scores are high and that relevant images may be visually diversified and not well clustered.

Topic models [3, 4] learn the latent topic of search results and rerank the results according to the probability that the result belongs to the topic. Some works combine clustering methods to filter out noises [6]. These methods work well for object-like queries, but may fail when relevant images are not aggregated in the top results.

Graph-based methods organize the search results using a graph with each node corresponding to an image and each edge connecting visually similar images. Then the original ranking is propagated through the graph [8, 9, 20]. The key problem of these methods is that irrelevant images that are originally ranked highly will stay highly ranked and be lowered down with little chance.

To deal with this problem, a spectral filter in [14] is used to detect the outliers before applying the graph-based method. In [18], a confident relevant sample detection method is proposed using sparsity and ranking constraints, followed by a kernel-ranking step. Though the two schemes have a good capacity of promoting relevant images and lowering down irrelevant images, they rely heavily on the assumption of visual aggregation of relevant images and may fail to detect confident samples in the case that relevant images in the search results are not well enough aggregated among top results.

All previous unsupervised methods merely use information limited in the search results under the current query. It is far from enough to improve search results when outliers are ranked highly or irrelevant samples visually dominate the top results rather than relevant images. The approach proposed in this paper makes use of helpful *deep contexts* to filter out irrelevant images, which handles the aforementioned issues. In contrast to supervised methods, our approach is automatic and needs no human annotations.

3. Contextual visual reranking

Given a query q, its search results are denoted as $\mathcal{I}_q = \{I_{q1}, \cdots, I_{qN}\}$. For convenience, we omit the subscript q and denote $\mathcal{I}_q = \{I_1, \cdots, I_N\}$. We use $\mathbf{z}$, a binary vector, to represent the outlier identification result. $z_i = 1$ means that image I_i is an outlier while $z_i = 0$ means that I_i is a relevant sample.

3.1. Outlier detection with deep contexts

Let's look at the example in Figure 2 that helps to illustrate the idea. In the bottom box are the top 10 images in

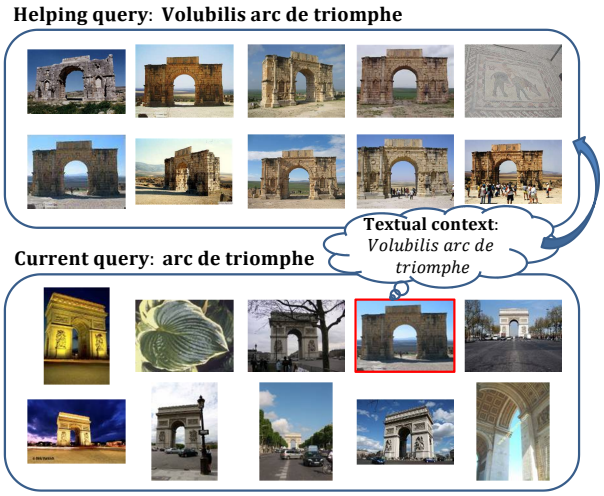

Figure 2. Outlier detection via deep contexts

the initial rank list for query "arc de triomphe". We aim to check if the fourth image highlighted in the red frame (denoted by I_o) is relevant. Though it is easy for human to recognize I_o as an irrelevant image, it is hard to distinguish it from the relevant images (the 3-rd and 5-th images) within the search results using their visual appearance. The images in the top box are the top 10 results for query "volubilis arc de triomphe". As can be seen, I_o is more similar to images in the top box and can easily be judged to belong to the top group rather than the bottom group. Thus, image I_o is likely to be irrelevant for query "arc de triomphe". In practice, this image is an outlier for this query.

This suggests a way to judge if an image I_k is relevant under the current query q by inspecting its relation to the search results from the queries (except the current query) formed by the textual context of this image. First, we make use of the textual context of I_k to form a set of queries $\mathcal{Q}_k = \{q_{k1}, \cdots, q_{k\bar{n}}\}$ ($q \notin \mathcal{Q}_k$) whose search results are called as *deep contexts*, which will be adopted to check the popularity of I_k to a query $\bar{q} \in \mathcal{Q}_k$. We observed that there may exist queries that are semantically equivalent to q. The text mining techniques, e.g., CYC [13] and WordNet [17], can be used to evaluate the synonymity between two queries to remove synonymous queries. But it is still not enough. Thus, we further remove more synonymous queries by checking the search results from $\bar{q}$ and q. To this end, we compute the synonymity between two queries,

$$\text{synonymity}(q, \bar{q}) = \frac{1}{N_q N_{\bar{q}}} \sum_{i=1}^{N_q} \sum_{j=1}^{N_{\bar{q}}} \text{sim}(I_{qi}, I_{\bar{q}j}), \quad (1)$$

where $\text{sim}(\cdot, \cdot)$ is the similarity function of two images, and $\text{synonymity}(q, \bar{q})$ measures the similarity of queries q and $\bar{q}$ from image search results of two queries. If the synonymity score is large enough, we regard it as a synonymous query

and remove it.

Now we turn to estimate the popularity $\text{popularity}(I, q)$ of image I to query q from the search results $\mathcal{I}_q = \{I_{q_1}, \cdots, I_{q_N}\}$ of q. It is observed that there are often multiple modes in the search results $\mathcal{I}_q$. Thus, it is not suitable to use all the images in $\mathcal{I}_q$ to check the popularity of image I to query q. Instead, we propose to estimate the popularity of image I in the local mode. To this end, we find the K nearest neighbors of I from $\mathcal{I}_q$ to form the neighborhood of I, which is denoted by $\mathcal{N}_q^I = \{I_{i_1}, I_{i_2}, \cdots, I_{i_K}\}$. The popularity is estimated as the average similarity between I and images in $\mathcal{N}_q^I$,

$$\text{popularity}(I, q) = \frac{1}{K} \sum\nolimits_{I_{\bar{q}} \in \mathcal{N}_q^I} \text{sim}(I, I_{\bar{q}}). \quad (2)$$

We compare the popularities of I_k to the query q and each $\bar{q} \in \mathcal{Q}_k$ satisfying that $\text{synonymity}(q, \bar{q}) < \delta$ to estimate the outlier score c_k that image I_k is an outlier for query q,

$$c_k = \max\nolimits_{\bar{q} \in \mathcal{Q}_k} f(I_k, q, \bar{q}), \quad (3)$$

where $f(I_k, q, \bar{q}) = 0$ if $g(I_k, q, \bar{q}) = \text{popularity}(I_k, \bar{q}) - \text{popularity}(I_k, q) \leqslant 0$ and $g(I_k, q, \bar{q})$ otherwise.

It is frequently observed that the outlier images are usually a minority in search results. In other words, the number of the outlier images is usually not too large. We make use of this property and formally formulate it as a sparsity regularization, $\|\mathbf{z}\|_1$. In summary, the objective function for outlier detection is formulated as follows,

$$\min\nolimits_{\mathbf{z}} \quad \lambda_1 \|\mathbf{z}\|_1 - \mathbf{c}^T \mathbf{z} \quad (4)$$

$$\text{s.t.} \quad \mathbf{z} \in \{0, 1\}^N, \quad (5)$$

where λ_1 is a constant to control the sparsity degree.

3.2. Outlier propagation

Though *deep contexts* help remove considerable amount of irrelevant images, there may still be some irrelevant images undetected due to the missing of useful textual context. Figure 3 shows such an example. Under query "race" there exist a group of images actually related to "jogging strollers" as shown in the figure. The image in the red box is successfully detected as irrelevant images because we find textual word "jogging strollers" in its textual context. However, for the image in the blue box, there are no useful words (e.g., no "jogging strollers") in its textual context. As a result, it cannot be judged as irrelevant images.

In some cases, e.g., the one shown in Figure 3, the irrelevant image (I_{31}) are visually similar to the outlier detected with deep contexts, which implies we can identify it through I_8. This suggests to propagate the outlier scores over the search results. To this end, we build a directed graph over

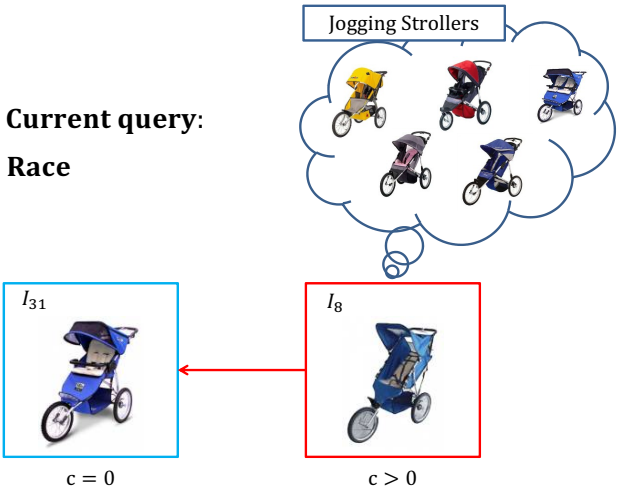

Figure 3. An example of image with useful context missed. I_8 indicates an irrelevant image that is detected, and I_{31} indicates another irrelevant image undetected because useful context is missed.

the search results. First, we build the visual graph by connecting all the pairs of images whose similarities are not small. And a weight $\text{sim}(I_i, I_j)$ is assigned to edge (i, j).

Then, we exploit *deep contexts* to reformulate the graph into a directed graph as our goal is to propagate the outlier scores. Considering two images I_i and I_j that are connected according to the visual similarity, denote $\mathcal{A} = \mathcal{Q}_i - \mathcal{Q}_j$, and $\mathcal{B} = \mathcal{Q}_j - \mathcal{Q}_i$. Define $\bar{q}_i^* = \arg\max_{\bar{q} \in \mathcal{Q}_i} f(I_i, q, \bar{q})$ and similarly define $\bar{q}_j^*$. We keep only the directed edge (i, j) if $\bar{q}_i^* \in \mathcal{A}$ and $c_i^* > c_j^*$; keep only the directed edge (j, i) if $\bar{q}_j^* \in \mathcal{B}$ and $c_i^* < c_j^*$; and build bidirectional edges (i, j) and (j, i) for I_i and I_j for other cases. The graph is represented as $\mathcal{G} = \{\mathcal{V}, \mathcal{E}, \mathcal{W}\}$ (for clarity, we drop the subscript q), where each node $\mathcal{V}$ corresponds to an image, $\mathcal{E} = \{(i, j)\}$, and $w(i, j) \in \mathcal{W}$ corresponds to the weight over (i, j), computed as $\text{sim}(I_i, I_j)$. For a directed edge (i, j), $\bar{q}_i^* \in \mathcal{A}$ means that $\bar{q}_i^*$ is a query that is missed in the textual context of I_j, so when $c_i^* > c_j^*$ and I_i is a confident outlier detected by *deep contexts*, the outlier score can be transmitted from I_i to I_j.

Given the detection result $\tilde{\mathbf{z}}$ using deep contexts, outlier propagation can be formulated as

$$\min\nolimits_{\mathbf{z}} \quad \sum\nolimits_{(i,j) \in \mathcal{E}} (z_i - z_j)^2 w_{ij} + \lambda_2 \|\mathbf{z}\|_1 \quad (6)$$

$$\text{s.t.} \quad z_i \geqslant z_j, \text{ if } (i, j) \in \mathcal{E}, (j, i) \notin \mathcal{E}, \quad (7)$$

$$z_i = \tilde{z}_i, \text{ if } \tilde{z}_i = 1, \quad (8)$$

where λ_2 is a constant to control the sparsity of outliers.

3.3. Overall objective function

We unite outlier detection through deep contexts and outlier propagation with visual similarities together to get an

overall objective function for outlier identification,

$$\min_{\mathbf{z}} \ \alpha \sum\nolimits_{(i,j)\in\mathcal{E}} (z_i - z_j)^2 w_{ij} + \beta\|\mathbf{z}\|_1 - \mathbf{c}^T\mathbf{z} \quad (9)$$

$$\text{s.t.} \ \ z_i \geqslant z_j, \ \text{if} \ (i,j) \in \mathcal{E}, (j,i) \notin \mathcal{E}. \quad (10)$$

Here, $\beta\|\mathbf{z}\|_1$ unites the $L1$ norm in Equation 4 and Equation 6. Furthermore, we rewrite the objective function in Equations 9 and 10 as a Markov random field form,

$$g(\mathbf{z}) = \sum\nolimits_i \phi_i(z_i) + \sum\nolimits_{(i,j)\in\mathcal{E}} \phi_{ij}(z_i, z_j), \quad (11)$$

where $\phi_i(z_i) = \beta - c_i$ and $\phi_{ij}(z_i, z_j) = \alpha w_{ij}(z_i - z_j)^2 + \infty\delta_{[z_i<z_j \& (i,j)\in\mathcal{E} \& (j,i)\notin\mathcal{E}]}$. As we can see, $\phi_{ij}(z_i = 0, z_j = 0) = \phi_{ij}(z_i = 1, z_j = 1) = 0$, and $\phi_{ij} \geqslant 0$. Naturally, the objective function $g(\mathbf{z})$ is submodular,

$$\phi_{ij}(0,0) + \phi_{ij}(1,1) \leqslant \phi_{ij}(0,1) + \phi_{ij}(1,0). \quad (12)$$

Thus, the problem can be solved by the graph cuts algorithm [10].

3.4. Graph reranking

We formulate the reranking problem as a semi-supervised learning problem. We regard the outlier identification result as the supervised information and propagate the information into the remaining search results. To this end, we build a graph over all the search results, where each node corresponds to an image and each edge is associated with a weight w_{ij} indicating the visual similarity between the corresponding images I_i and I_j. We denote this graph by a weight matrix $\mathbf{W}$. The objective function is written as follows,

$$\min_{\mathbf{r}} \ \ \|\mathbf{W}\mathbf{r} - \mathbf{r}\|_2 \quad (13)$$

$$\text{s.t.} \ \ r_i = \bar{r}_i, \ \text{if} \ i \leqslant m. \quad (14)$$

Here, $\mathbf{r}$ is the reranking score vector that will be used to reorder the search results, and $\bar{r}_i = h(z_i, \tilde{r}_i)$ with $\tilde{r}_i$ being the initial ranking of image I_i. m indicates the number of top images over which outlier detection and propagation has been conducted. In our implementation, we found $\tilde{r}_i = 2m - i$ and $h(z_i, \tilde{r}_i) = (1 - z_i)\tilde{r}_i$ works well. Suppose

$$\mathbf{W} = \begin{pmatrix} \mathbf{W}_{mm} & \mathbf{W}_{mn} \\ \mathbf{W}_{nm} & \mathbf{W}_{nn} \end{pmatrix},$$

where $n = N - m$ is the number of remaining images to be propagated. Then $\mathbf{r}_n = (\mathbf{I} - \mathbf{W}_{nn})^{-1}\mathbf{W}_{nm}\mathbf{r}_m$, where $[\mathbf{r}_m^T\mathbf{r}_n^T]^T = \mathbf{r}$.

3.5. Time complexity

Suppose that we conduct outlier detection and propagation over top m top images. The synonymity computation using deep contexts in Equation 1 takes $O(m^3)$, and the popularity computation in Equation 2 takes $O(m^2)$. The graph cuts algorithm solving Equation 11 takes $O(m^3)$. In our implementation, m is a small number that can guarantee the reliability of top images for outlier identification. For example, our reranking algorithm achieves the best performance on the INRIA dataset when $m = 40$. Therefore, outlier identification is very efficient. Because $\mathbf{W}$ is a sparse matrix the graph reranking takes $O(mN)$ to compute rank scores, where N is the length of the ranking list. On a desktop with 2.66GHz Intel Core Duo and 2GB memory, our MATLAB implementation takes about 0.6 seconds on average to rerank a query when m and N are 40 and 200.

4. Experiments

Experiments are conducted on the INRIA dataset [11] that includes 353 web queries and 71,478 web images. There are about 200 images for each query. Each image has an initial ranking score from a web search engine, a ground-truth label indicating whether it is relevant to the query, and a textual context associated with the image.

We use Stanford Named Entity Recognizer [5] to process textual words from the textual contexts for top 50 images in each query, and finally get 29,949 distinct textual words in total. Furthermore, we use these textual words as queries to search for images and crawl top 50 images for each query. Therefore, we obtain about 1.5 million images, organized for deep contexts.

We densely sample 128-dimensional SIFT descriptors from 16×16, 24×24 and 32×32 sized patches at every 8 pixels step from each image with VLFeat [19]. Then, a random collection of 200,000 descriptors is used to construct 1024 visual words by performing a k-means clustering. Following the scheme of spatial pyramid matching [12], we use 1×1, 2×2 and 3×1 sub-regions to compute a bag-of-words descriptor, a $d = 1024 \times 8$-dimensional feature $\mathbf{x}$, to represent each image. The similarity between two images I_1 and I_2 is evaluated as $\text{sim}(I_1, I_2) = \frac{\sum_{i=1}^{d}\min(x_{1i}, x_{2i})}{\sum_{i=1}^{d}\max(x_{1i}, x_{2i})}$. In evaluating the popularity in Equation 2, the neighborhood size is set as 5 in our implementation. In the objective function to identify outliers, we use $\alpha = 0.1$, $\beta = 0.006$. Experimental results show that when $\alpha \leqslant 0.1$, the scale of α and β makes little influence on the precision of outlier identification. However, the precision is sensitive to the proportion of α and β. $\alpha/\beta = 16$ yields the best performance.

In order to measure the reranking performances for each method, we adopt the Normalized Discounted Cumulated Gain (NDCG) and Mean Average Precision (MAP) which are standard evaluation criteria in information retrieval.

4.1. Performance comparison

We compare our approach with state-of-the-art reranking methods to demonstrate the effectiveness of our approach.

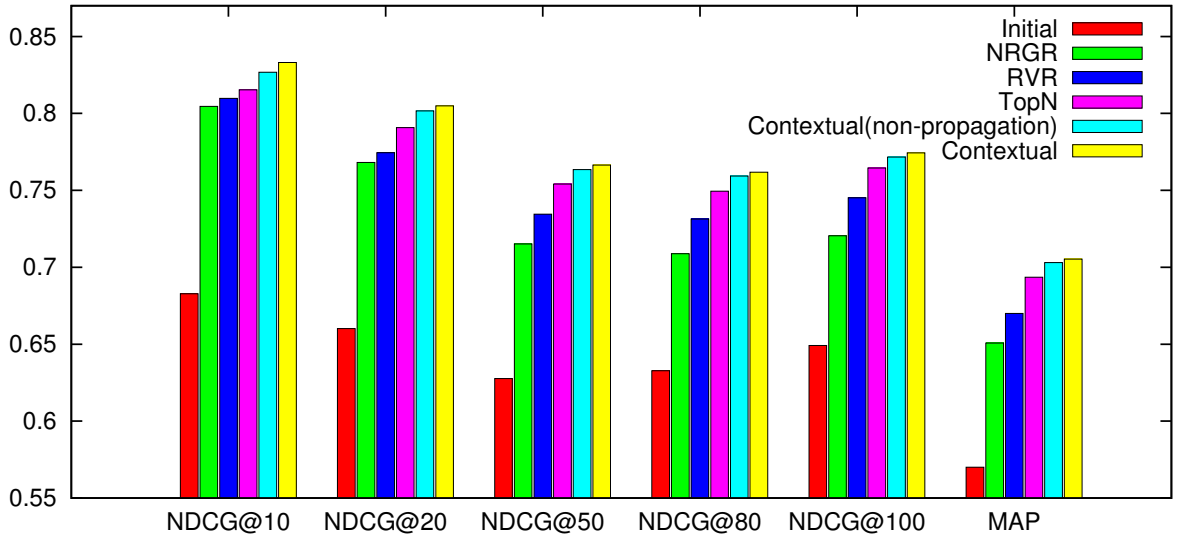

Figure 4. Experimental results on the INRIA dataset

Supervised reranking methods such as [7] are not used for comparison here, because these methods need training data while our methods do not.

- Robust Visual Reranking (RVR) [18] - A kernel-based reranking method using confident samples selected based on sparsity and ranking constraints.
- Noise Resistant Graph Ranking (NRGR) [14] - A graph-based reranking method using confident samples selected using Specfilter.
- TopN - The top N images for each query are assumed to be relevant images, and the kernel-based reranking [18] is performed for reranking. N is tuned in a brute-force manner to get the best performance.

Figure 4 shows that for the INRIA dataset, the performances of all reranking algorithms are much better than the initial rank, which demonstrates the effectiveness of doing search result refinement. The MAP of our method is about 0.135 higher than that of initial rank. The significant improvement from TopN to non-propagation contextual reranking demonstrates the effectiveness of our proposed outlier detection scheme with the use of *deep contexts*. Furthermore, the improvement from non-propagation contextual reranking to contextual reranking with propagation shows the effectiveness of outlier propagation. The ratio of effectiveness of outlier detection to outlier propagation is about $4:1$. Compared with RVR and NRGR, our MAP gets 0.035 and 0.055 improvements.

Some reranking results are listed in Figure 5. These queries have comparably very unreliable initial ranking results, which leads to little improvements or even worse performance using previous algorithms. However, our approach identifies the irrelevant images precisely and improves the reranking results significantly. For example, in the third subfigure, the current query is "Leeks". Our contextual visual reranking detects the irrelevant images that dominate in query "soup of leeks and potatoes" and query "leek quiche", and pulls the real "leek" images up. In the fourth subfigure with query "Jack Black", our approach identifies the irrelevant images related to gambling game "Blackjack" successfully and promotes the real positive images with actor "Jack Black".

4.2. Empirical analysis

4.2.1 Improvement by irrelevant images identification

In order to illustrate the effectiveness of our irrelevant images identification, we analyze the improvement of precision for relevant samples (remaining after outlier removal) by our approach. Figure 6 shows the results with the initial results as the baseline. As we can see, our method makes great improvement compared to initial results, which makes the subsequent reranking performance greatly improved. In particular, we analyze the precision at different top images level, because images with high initial ranks play a more important role than the others in the reranking phase. We can see that the precision for top 10 images improves from 0.668 to 0.731, which is very significant.

4.2.2 Different initial ranking score settings $\tilde{\mathbf{r}}$ and number of candidates for outlier identification m

Figure 7 shows that linearly decreasing initial ranking scores with $m = 40$ achieves the best performance in our

Figure 5. Visual results. The first row shows top-12 images ranked by the search engine, the second row ranked by Robust Visual Reranking, the third row ranked by Noise Resistant Graph Ranking, and the last row ranked by our approach.

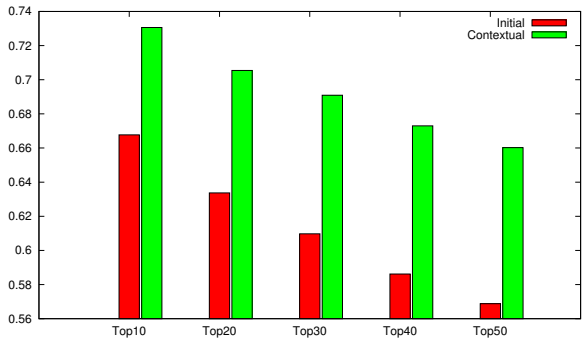

Figure 6. Precision for relevant images after removing outliers using our approach (contextual). *Initial* regards all top N images as relevant images.

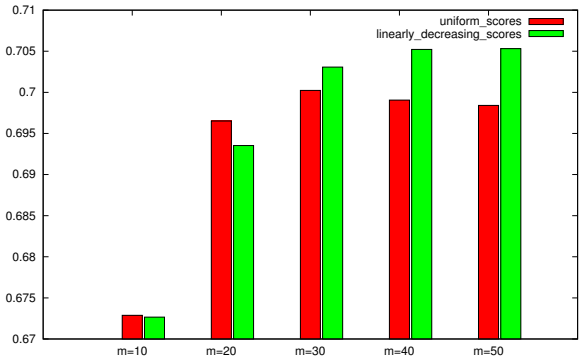

Figure 7. MAP under different initial ranking scores settings $\tilde{\mathbf{r}}$ and number of candidates for outlier identification m

approach, getting a 0.7053 MAP. This is because too few confident samples are not strong enough to represent the dominant visual features of a query, while too many confident samples may include too many noises and lose reliability. This also indicates that for the INRIA dataset, top 40 images have comparably high reliability.

With the increase of m, linearly decreasing initial ranking scores shows better performances than uniform initial ranking scores. This illustrates the importance of initial high-rank images. Within top 20 images, the importance of confident samples does not show great difference. While after top 20 images, the importance of confident samples decreases gradually. As a result, linearly decreasing initial ranking scores is a much better choice than uniform one for the INRIA dataset.

5. Conclusions

In this paper, we propose to exploit *deep contexts*, sets of search results returned from the queries formed by the textual context of one image, to detect irrelevant images and subsequently propagate the outlier scores to other images whose useful textual context is missed. We formulate the two strategy together and get an MRF formulation that is effectively solved by graph cuts. Then a graph reranking algorithm is used to reorder image search results. Unlike the previous methods whose performances are unsatisfactory when relevant images in search results are not visually aggregated or dominant, our approach is more robust and powerful. Experimental results show that our method performs significantly better than previous methods.

References

[1] N. Ben-Haim, B. Babenko, and S. Belongie. Improving web-based image search via content based clustering. In *CVPRW '06.*, page 106, june 2006. 2, 3

[2] T. Berg and A. Berg. Finding iconic images. In *CVPRW '09.*, pages 1 –8, june 2009. 2, 3

[3] R. Fergus, F.-F. Li, P. Perona, and A. Zisserman. Learning object categories from google's image search. In *ICCV*, pages 1816–1823, 2005. 2, 3

[4] R. Fergus, P. Perona, and A. Zisserman. A visual category filter for google images. In *ECCV (1)*, pages 242–256, 2004. 2, 3

[5] J. R. Finkel, T. Grenager, and C. D. Manning. Incorporating non-local information into information extraction systems by gibbs sampling. In *ACL*, 2005. 5

[6] M. Fritz and B. Schiele. Decomposition, discovery and detection of visual categories using topic models. In *CVPR*, 2008. 2, 3

[7] D. Grangier and S. Bengio. A discriminative kernel-based approach to rank images from text queries. *IEEE Trans. Pattern Anal. Mach. Intell.*, 30(8):1371–1384, 2008. 2, 6

[8] W. H. Hsu, L. S. Kennedy, and S.-F. Chang. Reranking methods for visual search. *IEEE MultiMedia*, 14(3):14–22, 2007. 1, 2, 3

[9] Y. Jing and S. Baluja. Visualrank: Applying pagerank to large-scale image search. *IEEE Trans. Pattern Anal. Mach. Intell.*, 30(11):1877–1890, 2008. 1, 2, 3

[10] V. Kolmogorov and R. Zabih. What energy functions can be minimized via graph cuts? *IEEE Trans. Pattern Anal. Mach. Intell.*, 26(2):147–159, 2004. 5

[11] J. Krapac, M. Allan, J. J. Verbeek, and F. Jurie. Improving web image search results using query-relative classifiers. In *CVPR*, pages 1094–1101, 2010. 2, 5

[12] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR (2)*, pages 2169–2178, 2006. 5

[13] D. B. Lenat. Cyc: A large-scale investment in knowledge infrastructure. *Commun. ACM*, 38(11):32–38, 1995. 3

[14] W. Liu, Y.-G. Jiang, J. Luo, and S.-F. Chang. Noise resistant graph ranking for improved web image search. In *CVPR*, pages 849–856, 2011. 1, 2, 3, 6

[15] Y. Liu, T. Mei, X.-S. Hua, J. Tang, X. Wu, and S. Li. Learning to video search rerank via pseudo preference feedback. In *ICME*, pages 297–300, 2008. 1, 2

[16] Y. Liu, T. Mei, M. Wang, X. Wu, and X.-S. Hua. Typicality-based visual search reranking. *IEEE Trans. Circuits Syst. Video Techn.*, 20(5):749–755, 2010. 2, 3

[17] G. A. Miller. Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41, 1995. 3

[18] N. Morioka and J. Wang. Robust visual reranking via sparsity and ranking constraints. In *ACM Multimedia*, pages 533–542, 2011. 1, 2, 3, 6

[19] A. Vedaldi and B. Fulkerson. Vlfeat: an open and portable library of computer vision algorithms. In *ACM Multimedia*, pages 1469–1472, 2010. 5

[20] J. Wang, Y.-G. Jiang, and S.-F. Chang. Label diagnosis through self tuning forweb image search. In *CVPR*, pages 1390–1397, 2009. 1, 2, 3

[21] R. Yan and A. G. Hauptmann. Co-retrieval: A boosted reranking approach for video retrieval. In *CIVR*, pages 60–69, 2004. 1, 2

[22] R. Yan, A. G. Hauptmann, and R. Jin. Multimedia search with pseudo-relevance feedback. In *CIVR*, pages 238–247, 2003. 1, 2

[23] L. Yang and A. Hanjalic. Supervised reranking for web image search. In *ACM Multimedia*, pages 183–192, 2010. 2