A SEMANTIC GRAPH-BASED ALGORITHM FOR IMAGE SEARCH RERANKING

Nan Zhao†, *Yuan Dong*†, *Hongliang Bai*‡, *Lezi Wang*†, *Chong Huang*†, *Shusheng Cen*†, *Jian Zhao*§

†Beijing University of Posts and Telecommunications,100876, P.R.China
‡France Telecom Research & Development - Beijing, 100190, P.R.China
§Patent Examination Cooperation Centers of the Patent Office, SIPO, Beijing, P.R.China
{nanzhao, yuandong, wanglezi, huangchong661100, censhusheng}@bupt.edu.cn
hongliang.bai@orange.com, zhaojian_2@sipo.gov.cn

ABSTRACT

Image search reranking has become a widely-used approach to significantly boost retrieval performance in the state-of-art content-based image retrieval system. Most of the methods merely rely on matching visual distances between query and initial results or among initial results to detect confident samples relevant to query. However, they may fail to rerank due to the existence of a huge gap between low-level visual features and high-level semantic concepts.

In this paper, we propose to detect reliable relevant samples based on a semantic image graph of labeled auxiliary dataset and Markov random walk algorithm. A graph-based rerank method is then presented to propagate the scores of detected confident samples to the rest. Our method is evaluated on the standard Paris dataset and a France dataset introduced by us. The performance is demonstrated to match or exceed the state-of-art.

Index Terms— Image search reranking, semantic graph, random walks

1. INTRODUCTION

Image retrieval and reranking have been one of attractive and challenging researches in the recent multimedia areas. Prevalent engineers of content-based image retrieval (such as Google Goggles) provide results relying on matching visual features of a query image to dataset. However, a large proportion of the initial search results is not relevant to the query image because a huge semantic gap exists between low-level visual features and high-level semantic concepts.

In order to offer a better user experience, the initial searching results should be reordered to improve the retrieval performance. The basic principle is to detect confident samples, considered as pseudo-positive samples, which will help rerank the rest of images. In the early work, top ranked images are directly taken as pseudo-positive samples to take part in the following rerank process [1, 2, 3]. However, these methods may not improve the performance or even worse it because false-positive samples always exist in the top list.

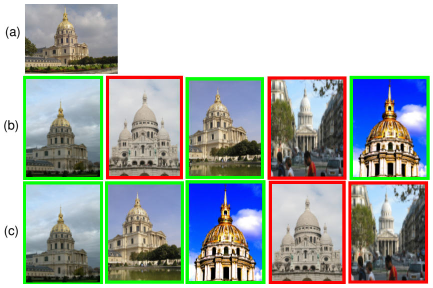


Fig. 1. (a) query image of Invalides in Paris. (b) The initial ranking results where green rectangular tags inliers and red for outliers. The second and the forth are Pantheon and Sacrecoeur in Pairs respectively. (c) Results after reranking.

In the recent researches, more reliable confident samples are learned to select instead of simple top N images of the initial list. One kind of reranking techniques tries to construct a new query from confident samples selected by robust spatial verification[4, 5, 6]. [4] uses a RANSAC-like geometric verification to remove false-positive samples and averages top confident samples to form a new query. In [5] min-hash is applied to detect noises in query images while [6] also learns models of noise features. The query region filtered out noises is taken as a new query to retrieval. The principle restriction of these methods is that they depend significantly on geometric verification, whose failure will lead to a collapse of query expansions. Another kind of reranking method selects pseudo positive samples based on the mutual visual relationships among top images in the initial ranking list [7, 8, 9]. Depending on the observation that outliers are less popular and more visually distinct than inliers, [8] introduces sparsity and ranking constraints to discover confident samples and reranks with kernel-based scheme. However, these methods may fail when the irrelevant images are uneasily distinguished from the relevant images visually.

An example is shown in Fig. 1 where irrelevant samples in the initial top ranking list have high visual similarity to the query, as well as relevant samples. In this case the previous methods may not work because they merely rely on comparing the low-level visual features distance to detect confident samples.

However, we observe that even though the visual distances between inliers and outliers are close, their tags or high-level concepts are obviously distinct, such as Invalides and Sacrecoeur. Inspired by [10] in which a semantic manifold is embedded to measure the image distance, we intend to merge semantic information to pick out the semantically relevant images and push down those semantically irrelevant.

In this paper, we propose to establish a semantic graph for auxiliary dataset of images, where every vertex is a labeled image and an edge links a pair of images who are semantically close. In [10] a semantic graph is established on ImageNet [11] dataset organized in a tree structure. Instead we treat each class independently due to the independence of concepts in our test dataset. Then top m initial results are utilized to select confident samples by a Markov random walk [12] inside the semantic graph. Finally, the rest of images are reranked based on the same graph where scores of detected confident samples are propagated. The performance of our algorithm is evaluated on standard Paris dataset, and France landmark dataset [13] which is crawled from Flickr, Bing and Google using queries of famous 78 France landmarks and 24 artworks in Louvre Museum.

The rest of this paper is organized as follows. Section 2 describes the process of building the semantic graph, detecting confident samples and reranking the rest relying on Markov random walk. In section 3 the performance of our algorithm is evaluated and compared with the state-of-art.

2. RERANKING

Our approach consists of two stages: off-line stage during which a semantic graph is built with labeled auxiliary data, and on-line stage in which confident samples are selected and reranking is applied.

We first define notations used in this paper. Let auxiliary dataset be $D = \{(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_N, y_N)\}$ of N images, where $\mathbf{x}_n$ is the low-level visual feature and y_n is the class label of image I_n. A weighted semantic graph is denoted by $G = (V, E, w)$, where V is the set of vertices and E is the set of edges with a weight function $w : E \to \mathbb{R}_+$. The matrix of graph G is defined as $\mathbf{W} \in \mathbb{R}^{N\times N}$, where w_{ij} represents a weight associated to edge $(i, j) \in E$. In addition, $\mathbf{h}^* \in \mathbb{R}^N$ is defined as an initial ranking vector, where h^*_i is non-zero if I_i belongs to initial ranking images set H or 0 otherwise. Similarly, we denote $\mathbf{h} \in \mathbb{R}^N$ as a confident samples vector, where the element is non-zero if its corresponding image is a confident sample. Let f be a function of detecting confident samples as:

$$\mathbf{h} = f(\mathbf{h}^*, \mathbf{W}) \tag{1}$$

Then the scores of detected reliable samples are propagated to rerank the rest images, giving the function

$$\mathbf{u} = g(\mathbf{h}, \mathbf{W}) \tag{2}$$

where $\mathbf{u} \in \mathbb{R}^N$ is a reranking vector and u_i is the score of image I_i.

2.1. Semantic Graph Building

The aim of building a graph is to connect the images which are visually as well as semantically related. Each vertex of G is a labeled image I in the auxiliary dataset, while each edge connects two of them undirectedly. A weight is assigned to each edge to reflect the similarity between the two vertices. Throughout the paper we may refer to the elements of V as the corresponding images.

Instead of linking every two vertices, we only connect one vertex to its k closest neighbors within the same class, which results in a sparse graph. For each vertex i, $K(i)$ is defined as a set of its k nearest neighbors whose class labels y equal to y_i. It is worth noting that the k closest neighbors include the target itself, representing self-transition. The similarity between V_i and V_j is computed through L1 distance between low-level visual features as:

$$s_{ij} = \frac{1}{1 + ||\mathbf{x}_i - \mathbf{x}_j||_1} \tag{3}$$

The edge weights are normalized starting from one node to its k connected neighbors so that rows of the graph matrix sum to 1. Therefore w_{ij} is defined as:

$$w_{ij} = \begin{cases} \dfrac{s_{ij}}{\sum_n s_{in}} & \text{if } j \in K(i) \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

In this way, each vertex is linked with those which share a high visual and semantic similarity. On the one hand, we only find the neighbors of the target node in its own class, which enables a semantic filtering and exploits a semantic coherence. On the other hand, for images that are highly close in semantic, visual feature distance guarantees a reliable measure for similarity.

2.2. Confident Samples Detection

The fundamental idea of this confident samples detection approach is that given a probable initial vertices distribution of a query, a final nodes distribution is found after a Markov random walking [12] inside the semantic graph.

Specifically, for a query image I_q, we firstly select top m similar images from the auxiliary dataset as candidate nodes set V_{h^*}. The similarity measure for initial ranking result relies on visual distance, so that an initial ranking vector $\mathbf{h}^*$ is defined as:

$$h_i^* = \begin{cases} \dfrac{s_{qi}}{\sum_n s_{qn}} & \text{if } n \in V_{h^*} \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

In order to find out nodes that are semantically close to the majority of V_{h^*}, Markov random walk is applied to pick out relevant nodes in dataset and suppress noises in V_{h^*}. According to [12], for an edge (i, j) one step transition probability from node i at time t to node j at time $t+1$ is

$$P_{t+1|t}(j|i) = \begin{cases} w_{ij} & \text{if } (i,j) \in E \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

The graph matrix $\mathbf{W}$ can be considered as Markov transition matrix. Since $\mathbf{W}$ has been normalized in terms of row, the probabilities of starting from vertex i to other vertices sum to one. Remembering that the k nearest neighbors include target vertex itself, we add self-transitions to the vertex so that w_{ii} is neither zero nor one.

Now given an initial vertices distribution $\mathbf{h}^*$ and a transition matrix $\mathbf{W}$, the reached vertices probability distribution $\mathbf{h}$ after one step of random walk is:

$$\mathbf{h}^{\mathbf{T}} = \mathbf{h}^{*\mathbf{T}}\mathbf{W} \tag{7}$$

We observe that relevant samples are more semantically aggregated in the top ranking list while classes that irrelevant samples belong to are usually more diverse. Due to this observation, the connected and nearby nodes will enhance each other and the disperse nodes will be suppressed through random walk.

After one step of random walk, new weights stored in $\mathbf{h}$ are assigned to each vertex $i \in V$, which are then sorted in a descending order. Intuitively top ranking nodes are more reliable than bottoms, therefore an adaptive threshold T is calculated to discard the bottom nodes as follows:

$$T = \frac{1 - \sum_{h_i > \alpha} h_i}{n} \tag{8}$$

where n is the number of vertices whose weights are less than α. If h_i is less than T, the corresponding vertex V_i is discarded, i.e. $h_i = 0$. It means that when the probabilities mainly distribute on the nodes with weights larger than α, the node whose probability is even less than the average distribution of the left nodes will be considered less reliable.

One run of a random walk is insufficient to recall enough confident samples. Therefore, the updated $\mathbf{h}$ is normalized again, considered as the positive feedback for the next random walk inside the semantic graph. The process is repeated r times so that more confident samples are selected and at the same time the order of them are reranked, illustrated in Algorithm 1.

In this way, images which are visually as well as semantically similar to the majority of the initial ranking list are selected as confident samples, and noises that have either diverse visual features or different class labels are removed.

Algorithm 1: Confident samples detection

```
input  : initial ranking vector $\mathbf{h}^*$, graph
         $G = (V, E, w)$, random walk times $r$
output: confident samples vector $\mathbf{h}$
$i := 0$
while $i < r$ do
    $\mathbf{h}^{\mathbf{T}} = \mathbf{h}^{*\mathbf{T}}\mathbf{W}$
    sort $\mathbf{h}$
    calculate $T$
    if $h_i < T$ then
        $h_i = 0$
    normalize $\mathbf{h}$
    $\mathbf{h}^* := \mathbf{h}$
    $i := i + 1$
```

2.3. Graph-based Reranking

The high-level idea of our reranking approach is to close the rank of two images whose path in our semantic graph is short. In terms of ranking vector this means that two similar images will have close weights.

Based on this intuition, we make use of the the confident vertices set V_h to solve the following objective function as:

$$\begin{aligned} &\mathbf{h}^{t+1} = \arg\min_{\mathbf{h}} \sum_{(i,j)\in E} (h_i^t - h_j^t)^2 w_{ij} \\ &\text{subject to } h_i^{t+1} = h_i^t \text{ if } i \in V_h, t = 1, 2, ..., l \end{aligned} \tag{9}$$

which is a convex optimization problem. Specifically, at time t the two vertices with higher similarity are forced to have similar weight at time $t+1$ to minimize the whole costs. Also, the constraint ensures that ranking vector will keep the weight of confident vertices. After each step V_h is updated. The process will be repeated l times and the final reranking vector $\mathbf{u} = \mathbf{h}^l$.

This problem could be solved by a simple system of linear equations based on random walk in [14]. Note that [14] requires the Laplacian matrix of G must be nonsingular. Since every vertex in our semantic graph connects to other $k - 1$ labeled images, the Laplacian matrix of G is nonsingular.

3. EXPERIMENTS

3.1. Experiment Setup

The performance of our approaches is evaluated on the standard Paris[13] dataset. Since it is relatively small including 6414 images of 11 landmarks, we establish a larger dataset called France. This dataset is crawled from Bing, Flickr and Google using queries for famous 78 France landmarks and 24 artworks in Louvre Museum, such as Amphi Theater and Mona Lisa, including 102 classes of 86717 images shown in

Fig. 2. Randomly selected images from the France dataset

Fig. 2. We also select the same 55 queries of Paris dataset as the queries of France dataset. As for low-level visual features, Harris-Laplace detectors [15] and SIFT descriptors [16] are used to describe the local visual information. Images are represented as bag of words with standard tf-idf scheme. We apply vocabulary tree [17] to train a vocabulary of 10^6 visual words. In the experiments, the following parameters are used: number of connected vertices for each target node $k = 10$, number of top nodes initially selected for confident samples detection $m = 10$, and $\alpha = 0.01$.

Besides, we use Average Precision to evaluate the performance of each query which is the area under the precision-recall curve. A mean AP (mAP) is calculated by averaging APs of all queries to assess a dataset.

3.2. Graph Construction and Parameters Selection

Firstly, for Paris dataset we simply use the standard offered tags of images to build the semantic graph, while for France dataset images are downloaded with tags.

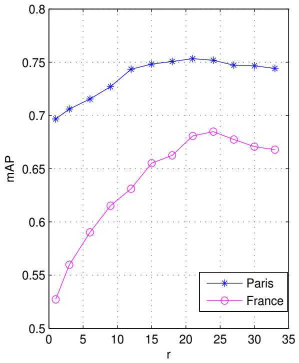
Parameter variation: We compare the performance of different r and l. The mAP is calculated with different r by setting l a constant, as shown in Fig. 3. The performance is relatively high when r falls between 19 and 24. In the following experiment shown in Fig. 4, we set $r = 20$ and vary l. The mAP tends to be steady when r is greater than 12. Therefore we select $r = 20$ and $l = 14$ and the mAP reaches 0.786 and 0.723 on Paris and France dataset respectively.

Graph Re-construction: Although the mAP for Paris dataset is relatively high (0.786) when tested on the previous semantic graph, a few queries including Eiffel-3 and Trimphe-5 fail. One reason for the failure is that a class called General contains images with various tags belonging to different landmarks, however, we take it as an independent class when building our graph. Another reason is that there is some mislabeled images, such as Trimphe-5 image labeled as Defense. Therefore, we relabeled the images in class General and build a new semantic graph. The mAP is improved to 0.826.

3.3. Performance Comparison

We compare our approach with the state-of-art reranking algorithms running on the Paris dataset.

a. Top N reranking: a baseline where top N images are selected as confident samples and relation graph is built only based on visual similarity.

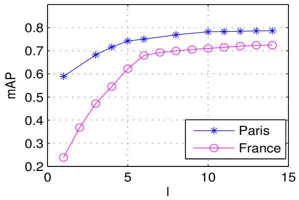


(a)

r	Paris	France
1	0.697	0.527
3	0.706	0.560
6	0.715	0.590
9	0.727	0.615
12	0.743	0.631
15	0.748	0.655
18	0.750	0.662
21	0.753	0.680
24	0.751	0.684
27	0.747	0.677
30	0.747	0.670
33	0.744	0.668

(b)

Fig. 3. Selection of r on Paris and France dataset. $l = 6$.



(a)

l	Paris	France
1	0.589	0.238
5	0.742	0.622
10	0.782	0.710
12	0.784	0.719
14	0.786	0.723
15	0.786	0.724

(b)

Fig. 4. Selection of l on Paris and France dataset. $r = 20$.

b. K-reciprocal Nearest Neighbors (KRNN): proposed by D.Qin *et al.* [18] where different similarity measures are used for different parts of initial ranking list.

c. Total recall II (TR II): proposed by O.Chum *et al.* [6] where the confident samples are selected through confuser filtering and then applied to incremental spatial reranking.

The retrieval performance of each algorithm on Paris dataset is displayed in Table 1. It shows that our approach have reached the state-of-the-art results, outperforming the KRNN and TR II 2.3% and 2.1% respectively on their best results in previous publication.

Table 1. Performance Comparison on Paris Dataset

method	Top N	KRNN	TR II	Ours
mAP	0.612	0.803	0.805	**0.826**

4. ACKNOWLEDGEMENT

The work is sponsored by co-research project SEV01100474 between France Telecom R&D Lab Beijing and Beijing University of Posts and Telecommunications, and Graduate Innovation funding of SICE, BUPT 2011.

5. REFERENCES

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

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

[3] J. Wang, Y.-G. Jiange, and S.-F. Chang, "Label diagosis through self tuning forweb image search," *Proc. CVPR*, pp. 1390–1397, 2009.

[4] O. Chum, J. Philbin, J. Sivic, M. Isard, and A. Zisserman, "Total recall: Automatic query expansion with a generative feature model for object retrieval," *Proc. ICCV*, 2007.

[5] O. Chum and J. Matas, "Unsupervised discovery of co-occurence in sparce high dimensional data," *Proc. CVPR*, 2010.

[6] A. Mikulik, M. Perdoch, and J. Matas, "Total recall ii: Query expansion revisited," *Proc. CVPR*, 2011.

[7] W. Liu, Y.-G. Jiamg, J. Luo, and S.-F. Chang, "Noise resistant graph ranking for improved web image search," *Proc. CVPR*, pp. 849–856, 2011.

[8] N. Morioka and J. Wang, "Robust visual reranking via sparsity and ranking constraints," *ACM Multimedia*, pp. 533–542, 2011.

[9] G. Park, Y. Baek, and H.Lee, "Re-ranking algorithm using post-retrieval clustering for content-based image retrieval," *Information Processing and Management*, vol. 41(2), pp. 177–194, 2005.

[10] Fang C and L. Torresani, "Measuring image distances via embedding in a semantic manifold," *Proc. ECCV*, 2012.

[11] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and Feifei Li, "Imagenet: a large-scale hierarchical image database," *Proc. CVPR*, 2009.

[12] N. Craswell and M. Szummer, "Random walks on the click graph," *Proc. SIGIR*, pp. 239–246, July 2007.

[13] J. Philbin, O. Chum, M.Isard, J. Sivic, and A. Zisserman, "Lost in quantization: Improving particular object retrieval in large scale image databases," *Proc. CVPR*, 2008.

[14] S. Rota Bulo, M. Rabbi, and M. Pelillo, "Content-based image retrieval with relevance feedback using random walks," *Pattern Recognition*, vol. 44, pp. 2109–2122, 2011.

[15] K. Mikolajczyk and C. Schmid, "Scale & affine invariant interest point detectors," *IJCV*, vol. 1(60), pp. 63–86, 2004.

[16] D. Lowe, "Distinctive image features from scale-invariant keypoints," *IJCV*, vol. 60(2), pp. 91–110, 2004.

[17] D. Nister and H. Stewenius, "Scalable recognition with a vocabulary tree," *Proc. CVPR*, 2006.

[18] D. Qin, S. Gammeter, T. Quack, and L. Van Gool, "Hello neighbors: accurate object retrieval with k-reciprocal nearest neighbors," *Proc. CVPR*, 2011.