Hybrid-Indexing Multi-type Features for Large-scale Image Search

†Qingjun Luo,‡Shiliang Zhang,†Tiejun Huang,†Wen Gao,‡Qi Tian

†School of Electronics Engineering and Computer Science, Peking University
‡Department of Computer Science, University of Texas at San Antonio

Abstract. Indexing local features with a vocabulary tree and indexing holistic features by compact hashing codes are two successful but separated lines of research. Both of the two indexing models are suited for specific features and are limited to certain scenarios like partial-duplicate search and similar image search, respectively. To conquer such limitations, we propose a novel hybrid-indexing strategy, which incorporates multiple similarity metrics into one inverted index file during off-line indexing. Hybrid-Indexing only requires the Bag-of-visual Words (BoWs) model as input for online query, but could obtain more satisfying retrieval results because the index file conveys hybrid similarities among images. Moreover, hybrid-indexing does not degrade the efficiency of classic BoWs based image search. Experiments on several public datasets manifest the effectiveness and efficiency of our proposed method.

1 Introduction

Though many successful image retrieval methods have been proposed in recent years, the existence of semantic gap still hinders the improvements of retrieval accuracy [1]. The major challenges come from two aspects, 1) it is not easy to capture the essential characteristic of an image by a single feature, if not impossible, and 2) it is hard to depict the actual search intention of the user.

Efforts to bridge the lack of coincidence between the extracted image information cues and the interpretation of human users are made in various ways. To obtain intrinsic descriptions of an image, different types of image features have been proposed. Local features and holistic features are two main categories. Not withdrawing their success, it has been more and more noticed that there is no a single type of feature which is optimal in all cases. For examples, local features with vocabulary trees are well suited to find partial duplicate objects, while global semantic attribute features aim to locate images that share similar semantic meanings. Different cues delineate distinct aspects of an image, and meet the users' various search intentions separately. Feature fusion is a reasonable option to leverage multiple cues [2, 3]. Unfortunately, incompatibility issues arise accordingly because of the modality difference of the fused features. In the meanwhile, extracting various types of features during online retrieval stage would greatly increase the query time.

Another line of feasible methods to incorporate different cues is combining multiple retrieval results during online retrieval [4–7]. However, these methods suffer from either the difficulties of measuring and combining dramatically diverse features, or the

computational expense which is introduced by online multiple feature extraction and fusion.

To the best of our knowledge, the effort on fusing different features or similarity metrics during the offline indexing is still limited. One of the original works is [8], which propose a semantic-aware co-indexing algorithm to jointly embed semantic attribute and local features in the inverted indices. [9] packs semantic relevant images into an uniform unit, *i.e.,* superimage, and indexes these semantically compact units instead of single images, thus largely reduced the memory consumption while obtaining both semantically and visually relevant images at the same time.

As an important procedure in BoWs based image retrieval systems, off-line indexing organizes images sharing a common visual word together into one image list, *i.e.*, inverted list, which could be accessed by the ID of this visual word. The performance of BoWs based image retrieval may be affected by two factors, 1) feature detection failure, which fails to extract accurate local features, and 2) quantization error which assigns non-relevant local features into one visual word. This leads to two possible flaws of the reverted index, respectively, *i.e*, the image list associating with a certain visual word misses entities that contains this visual word, and it also may contains images that are actually non-relevant.

In the meanwhile, images in the list are assumed to be visually relevant, but their overall distribution remains uninvestigated. Obviously, images that are not only visually similar but also share certain holistic consistencies should be reasonably organized together. That motivates us introduce holistic cues into the visually relevant image list.

In this paper, we introduce hybrid-indexing as a novel index fusion algorithm. Instead of extracting multiple image features during online retrieval or only consider two kinds of information cues such as co-indexing [8] or sharper image [9], our method incorporates multiple cues *simultaneously* into the inverted index. We investigate the consensus of images in various aspects and re-organize the index structure. The procedure only takes place off-line, and does not sacrifice online retrieval efficiency. Specifically, for each image lists that associating with a certain visual word, we build several directed graphs according to different holistic neighboring relationships. Then multiple graphs built by different cues are fused together by consolidating edges. A link analysis on the resulting graph is conducted to obtain the PageRank vector, which can be considered as a measurement of the image consistencies to each other. We remove the *isolated* images that are dissimilar to the others, and for the *significant* images with higher PageRank values, we replace them with their affiliating superimages [9]. The generated hybrid-indexing serves as the updated inverted index file, and can be accessed with the classic BoWs based retrieval model without any further modifications. The flowchart of the proposed method is illustrated in Fig. 1.

The main contribution of the proposed approach can be summarized as follows. 1) To our best knowledge, this is one of the few works on fusing different cues during off-line indexing period. 2) The proposed framework does not limit the quantity of fused cues and allow simultaneous consideration of local features and multiple holistic features, which may be infeasible in former works. 3) Our method does not need to extract multiple features during online retrieval, but manages to obtain consistent images not only are partially duplicated but also share a similar holistic characteristic.

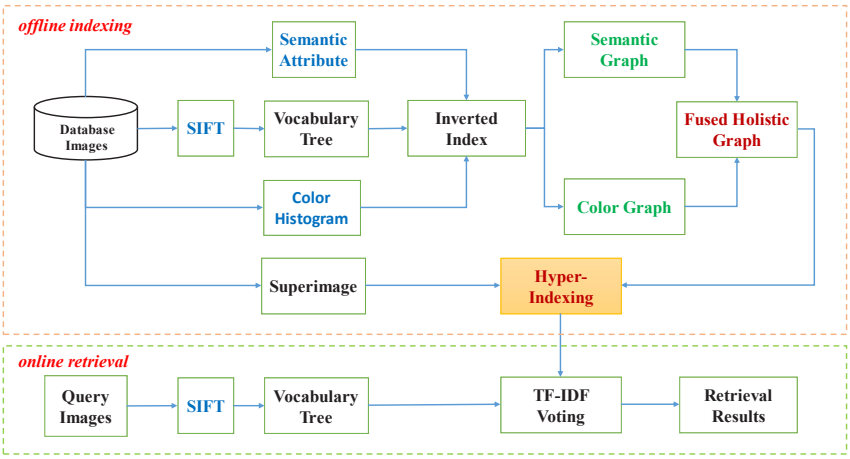

Fig. 1. Framework of the proposed method.

2 Related Work

There are two major categories of scalable image retrieval algorithms: searching for exact or near-duplicated images by indexing local features with vocabulary trees, and hashing holistic image features to binary codes to find similar images. We briefly review the two lines of retrieval strategies and the efforts have been made to combine them.

Local Features with Vocabulary Trees. Conventional BoWs model based image retrieval systems have been proven to be a great success in the past few years. By quantizing [10] local invariant image features [11–14] with vocabulary trees [15], an image is converted into a BoWs vector. The vocabulary tree is usually hierarchically trained on lots of local descriptors, and each of the leaf nodes is considered as a visual word. BoWs based retrieval systems commonly employ inverted index files to organize images [15], and in this file, each visual word is associated with a list of images containing this visual word. Retrieved candidates are further improved by spatial verification [16, 17], query expansion [18], hamming embedded [19], high-order features [20]. Because the images are essentially described by local features, BoWs model is well suited for near-duplicate image search, while the results may show less global consensus due to lack of holistic depictions.

Holistic Features with Hashing Methods. Unlike local features, holistic features such as color histogram and GIST [21], aim to delineate overall distributions of an image, hence could capture certain aspects of the image global characteristics. Recent developments in large scale image recognition and classification also contribute to provide semantic-aware image features [22, 23]. The outcomes of multi-class classifiers, often

referred as semantic attributes [22–24], present a strong cue to find semantically relevant images. Holistic features are often indexed by locality sensitive hashing [25], and the resulting hash codes can be efficiently compared using Hamming distances. Since holistic features are not as invariant as local features, thus the retrieved candidates often miss near-duplicate objects.

Local Feature and Holistic Features based Retrieval Fusion. Near-duplicate image retrieval using local features and similar image retrieval with holistic features are two lines of research. Although the limitation of individual retrieval schema is obvious, efforts made to fuse multiple features are rare. Combining local and holistic features can be conducted on the early feature fusion stage [2] or the later retrieved results fusion stage [4, 6]. However, for image feature fusion, it is difficult to leverage various modalities of the fused feature. When it comes to ranking level fusion, multiple feature extraction and online re-ranking computation often introduce dramatic query time increasement. For example, [6] conducts query specific fusion during online retrieval and achieves a decent precision while consuming much memory overhead and query time. Works to fuse local features and holistic features in off-line indexing stage is either not optimal in identifying isolated images [8] or only could deal with one holistic feature at a time [9]. Our work shares some common properties with [8]. The differences are we involve more features and propose a more principled ranking strategy to spot and delete isolated images. Compared with [8], we achieve a more significant improvement over the baseline BoWs model.

3 Proposed Approach

We propose to fuse multiple image cues to build hybrid-indexing which embraces both local and multiple holistic features simultaneously. Various holistic features are extracted and superimages are constructed firstly, which is described in Sec. 3.1. In an inverted index file, images containing a common visual word form a short image list. We hence perform the following steps for each of the image lists throughout the whole index structure. For each of such lists, according to the kNN(k-Nearest Neighbor) relationships in every individual holistic feature space which is obtained in Sec. 3.1, multiple directed graph are built up, then these graphs are fused by consolidating edges to obtain a final graph (Sec. 3.2). PageRank values are computed on this graph to rank these images. For images with low rank values, we identified them as *isolated* images and remove them from the image list. For the *significant* images with high rank values, we replace them with their affiliating superimages [9]. This procedure is described in detail in Sec. 3.3 and 3.4.

3.1 Holistic Feature Extraction and Superimage Generation

The proposed method is not restricted to the usage of the type and quantity of holistic features. In our current implementation, we employ two commonly used holistic features, *i.e.*, semantic attributes and color histograms.

Semantic attributes are commonly computed as the classification scores of object classifiers. We follow the method of [8] to learn 1000 object SVM classifiers from the training images in LSVRC 10[26], which is a subset of ImageNet dataset. Dense HOG and LBP features are extracted and further encoded by local coordinate coding. The margin scores of these SVMs are used as semantic attribute features. Our test sets are independent with ImageNet, hence we do not implicitly assume the query or the dataset are related to one object category in these semantic attributes. Distance of the 1000-dimensional features are measured by *cosine* distances, a common distance metric for floating-point features.

Colors are important visual cues to represent the image content. We employ distribution statistics from the HSV color space as image descriptors. Following the method of [3], we only focus on the saliency regions instead of extracting features from the whole image. Spectral Residual Model (SRM) is employed to automatically extract saliency regions, which are independent from any prior knowledge. The resulting color descriptor shows higher robustness than considering the whole image [3]. The dimension of the HSV color histogram is 48. We also use *cosine* distance to measure the distances between color features.

We follow exactly the same experimental and parameter settings of [9] to generate superimages. Specifically, mutual-*k*NN graph is built first based on semantic distances between images. We employ semantic attributes to build these sparse graphs and Euclidean distances to measure their distances. Maximal cliques searching algorithm is then employed to generate superimage candicates, followed by a greedy ranking algorithm to rank and select the final superimages. Each superimage contains a single or multiple images which are semantically relevant to each other, and thus can be considered as a representation of a particular semantic meaning. The superimage generation process can be completed off-line efficiently. For each image, we maintain an index structure to record its affiliating superimage, *i.e.*, the superimage that contain it. Note that, the superimage index structure will be used only in the procedure of hybrid-indexing construction (Sec. 3.4), and is no need for online retrieval stage.

3.2 Graph Generation and Fusion

The inverted index files organize the corresponding relationship between visual words and images. A set of images containing a certain visual word can be regarded as a set of locally visual relevant images. However, the discriminative capacity of a single local descriptor is limited, and the consistency of images within the set may not be reliable enough consequently. Therefore, we turn to further investigate their relationships in holistic feature spaces. Based on the obtained multiple holistic features, we build several relationship graphs over them respectively. We define the image set I_v associating with the visual word v as:

$$I_v = \{i | i \in D, v \in i\} \tag{1}$$

where D denotes the whole image dataset. For each image i in I_v, we link it with its *k*NN(k-Nearest Neighbor), then we use I_v as the vertex set and the directed connections

as edge set, a graph

$$G =< I_v, E >$$

depicting their relationships is obtained, where

$$E = \{< i, j >) | i \in kNN(j), i, j \in I_v\} \tag{2}$$

in which i and j are two images in the list and $kNN(j)$ denotes *k*NN set of image j.

Based on different types of holistic features, a series of graphs $G^h =< I_v, E^h >$ can be generated. To utilize the advantage of multiple holistic cues simultaneously, we fuse them together into one graph $G =< I_v, E >$ with the same vertex set I_v and $E = \bigcup_h E^h$. That means if image i is *k*NN of j in any holistic feature space, there is a directed link pointing from i to j in the fused graph. Links that occur in more than one individual holistic graph imply much reliable relationship between images, and they will be favored in the following link analysis step. Note that, the procedure does not restrict the type of fused holistic feature, or the quantity of holistic cues. If there is only one holistic feature, the fused graph is identical to the one which is built on it.

An illustration of the process of graph fusion is shown in Fig. 2.

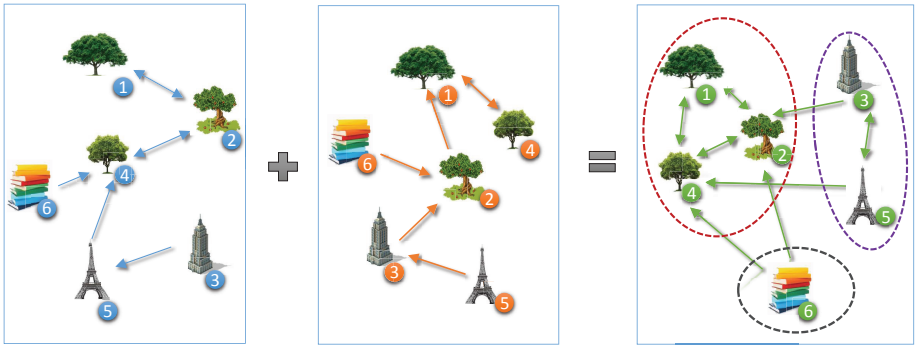

Fig. 2. An illustration of the process of multiple holistic graphs fusion. The toy example demonstrates two holistic graphs built over 6 images. A directed arrow represents that the end point is the *k*NN of the starting point ($k = 1$ in this example). The fused graph is constructed over the same vertex set and consolidated edges. It is can be observed that the fused graph captures more meaningful holistic distribution characteristic. PageRank algorithm is conducted on this fused graph to compute the rank value vector, and obtain the result: [0.2824, 0.3028, 0.0435, 0.3028, 0.0435, 0.0250]. Based the rank values, image 1, 2, and 4 are identified as *significant* images (red dotted circle), and image 6 is regarded as *isolated* image (black dotted circle).

3.3 Significant and Isolated Image Detection

Given a graph G either generated by a single holistic feature or fused by multiple relationship graphs, the connectivity of a node reflects its global consensus to the others.

This motivates us to conduct a link analysis [27] on graph G to rank according to their node connectivity. PageRank is a commonly used ranking algorithm to weight the importance of nodes in a graph. Because the graph G is built according to their neighboring relationships, a node is more important or relevant if it is more likely to be visited. A $|I_v| \times |I_v|$ connection matrix M is defined as $M_{ij} = 1/\deg(i)$, if $< i, j > \in E$, and $M_{ij} = 0$ otherwise, where $deg(i)$ denotes degree of node i. Normally, PageRank algorithm adopts a damping factor to guarantee its convergence. We empirically set it as *0.85* in all experiments.

After several iteration steps, a PageRank vector is computed and each image is assigned with a rank value, which depicts its importance in the graph G. To a certain extent, the obtained rank value can be deemed as the confidence of an image is relevant to this visual word. The images are ranked according to their PageRank values. Images with high values are considered as *significant* images, while the ones with low values are regarded as *isolated* images. Specifically, denote $PR(i)$ as the rank value of image i, α and β as the upper and lower threshold respectively, *significant* image set $\hat{I}_s$ is defined as

$$\hat{I}_s = \{i | i \in I_v, PR(i) > \alpha/|I_v|\} \tag{3}$$

and the *isolated* image set $\hat{I}_i$ is defined as

$$\hat{I}_i = \{j | j \in I_v, PR(j) < \beta/|I_v|\}. \tag{4}$$

In this way, images within the list I_v are divided into three categories, *i.e.*, *significant* images, *isolated* images, and the rest images. In the process of hybrid-indexing construction, images in different categories will be treated differently.

3.4 Hybrid-indexing

Among all the images containing the visual word v, *significant* images are trusted to contain this visual word with certainty, and although v is also extracted on the *isolated* images, it is more likely that they may influenced by image noises or quantization errors. This motivates us to remove *isolated* images from the image set I_v and replace *significant* images with their affiliating superimages.

The concept of superimage is proposed in [9]. It consist of single or multiple compact semantic relevant images, and each image within is the mutual-*k*NN of each other. Once there is a *significant* image located, we replace it with its affiliating superimage. We maintain an index structure which is built in advance (Sec. 3.1) to record the superimage membership information. During the replacement process, the introduced superimages are unpacked into its affiliating images, thus for a *significant* i, the process is also equivalent to inserting images that share the same membership. The inserted images are assigned Term Frequency (TF) the same as image i. Note that, the size of superimages could be equal or greater than one. If the superimage size of i is one, in which case the superimage of i is identical to itself, image i in I_v remains unchanged.

Apart from *significant* images and *isolated* images, the rest of the images also keep unchanged during the procedure.

Specifically, a line of the constructed hybrid-indexing is obtained by

$$I_v{}^{new} = I_v - \hat{I}_i - \hat{I}_s + \{SI(i)|i \in \hat{I}_s\} \tag{5}$$

where $SI(i)$ is the affiliating superimage of i.

After the removal and replacement procedure, the image list I_v is updated by a new set $I_v{}^{new}$. Process all image lists in the inverted index file by order, hybrid-indexing with multi-type cues is generated. Image retrieval with the renewed hybrid-indexing does not need to alter the local feature extraction or the online retrieval stage, and it can be assembled with conventional BoWs model directly.

4 Experiments

4.1 Experimental Setup

Three different image retrieval tasks are conducted to evaluate the proposed method, *i.e.*, object search on UKbench [15], scene image search on INRIA Holidays [19], and large-scale image search on a dataset built by mixing MIRFLICKR-1M [28] collected from Flickr[1] with UKbench.

UKbench dataset contains 2,550 objects under 4 different viewpoints and illuminations. The retrieval performance is measured by the recall of top-4 retrieved images which is referred as N-S score. N-S score ranges from 0 to 4, indicting none or all of the relevant images are returned. INRIA holidays dataset includes 1,491 annotated personal holiday photos and 500 of them are served as queries. Performance on this dataset is measured by mAP (mean Average Precision). For large-scale image search, we use the public large dataset MIRFLICKR-1M [28] consisting 1 million real-world images as distractors, and mix it with UKbench dataset. Images from UKbench are employed as queries and N-S score is adopted as the performance metric.

4.2 Image Search

In this section, We test the proposed method with object search on UKbench and scene image search on INRIA Holidays. Different vocabulary trees are utilized to test if our method is sensitive to the vocabulary tree structures, which usually result in different levels of quantization errors in classic BoWs based image search. We train a visual vocabulary tree with branch number B = 10 and layer number L = 5 (denoted as $T10^5$). It is trained against a separate data set consisting of 50,000 images randomly selected from ImageNet dataset. Besides of that, we also utilize the local features and the pre-computed visual words provided by the original authors of the two datasets for fair comparison. Despite of the authors of INRIA Holidays provide several varieties of vocabulary trees, for both of the two datasets, we employ vocabulary trees with B = 10 and

Table 1. The retrieval performances of baseline methods

Dataset	UKBench(N-S Score)		Holidays(mAP)	
Performance	$T10^5$	$T10^6$	$T10^5$	$T10^6$
	2.8175	3.1664	58.3%	66.5%

L = 6 (denoted as $T10^6$) respectively. Retrieval performances of baseline approaches are summarized in Table 1.

Three parameters need to be decided during the process of constructing hybrid-indexing, *i.e.*, count of neighboring images k in Eq. 2, lower threshold α in Eq. 3, and upper threshold β in Eq. 4. We tune these three parameters on UKbench with $T10^6$, then apply them to the rest of our experiments.

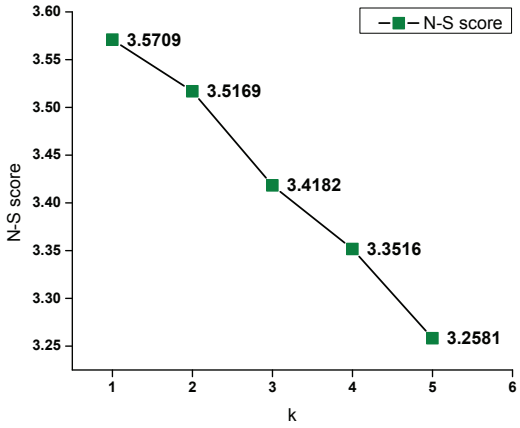

Fig. 3. The influence of the parameter k. Values larger than 1 yield less competitive results, thus we just search the nearest neighbor when constructing holistic feature graph.

We first fix α as 1 and β as 0, in which case no images will be identified as *significant* or *isolated* images, and investigate the influence of the count of neighboring images k in Eq. 2. Each image in Eq. 2 is linked to its k nearest neighbor, so the total count of edges in graph G is computed as $|E| = k * |I_v|$. That means larger k introduces more connections and makes the graph denser. We demonstrate the influence of the parameter k on UKbench in Fig. 3. Interestingly, values larger than 1 yield less competitive precision than setting k as 1, *i.e.*, only searching for the nearest neighborhood. The reason might be connecting excessive edges would introduce less stable relation-

[1] http://press.liacs.nl/mirflickr/

ships, thus is harmful to the performance. According to the tuning result, we set the count of neighboring images $k = 1$.

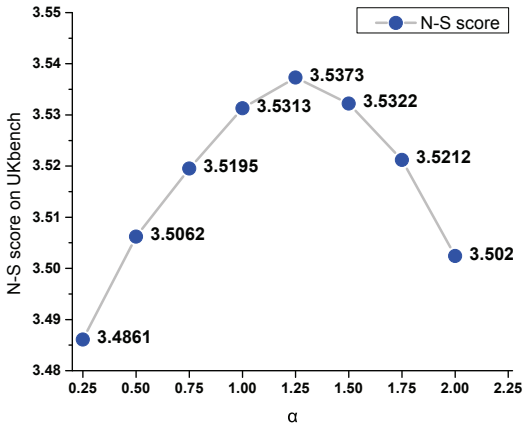

Fig. 4. The influence of the parameter α controlling the choice of *significant* images. It is clear that replacing *significant* images with their affiliating superimage improve the retrieval precision dramatically. With larger α, N-S score keeps increasing until a performance drop occurs.

Next, we keep setting β as 0, in which case the algorithm removes no *isolated* images and observe the impact of the choice of upper threshold α controlling the *significant* image selection. The experimental results are shown in Fig. 4. It can be seen that our algorithm shows significant improvement over the baseline (3.1664 of $T10^6$). The performance keeps increasing with reasonably lager α, indicating the benefit of embracing superimages is obvious. However, after a certain value, a performance drop occurs if we keep increasing α. The reason might be because larger α yield less *significant* images, thus not fully utilize the potential of the superimage replacement procedure. According to Fig. 4, in all of the following experiments, we set the upper threshold $\alpha = 1.25$.

After that, we fix the upper threshold and tune the lower threshold, *i.e.*, β. The influence of the parameter β is demonstrated in Fig. 5. It proves the advantage of the procedure of isolated removal, although one can see that it does not help as much as superimage replacement. With increasing β, the algorithm removes more and more *isolated* images, and results in more compact index file and less memory consumption. The performance drops when excessive images are wrongly recognized as *isolated* images, and the precision is hindered consequently. Based on this observation, β is fixed as 0.25 in the following experiments.

Note that, in our current implementation, we adopt a simple parameter tuning strategy. Carefully evaluating the combinations of k, α and β would obtain better parameter

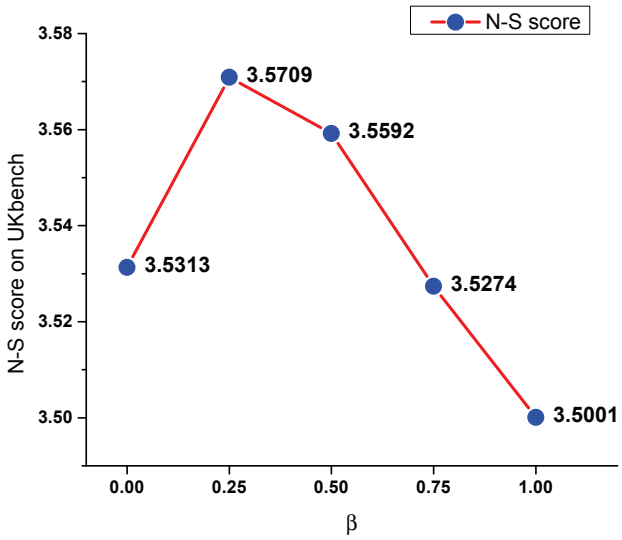

Fig. 5. The influence of the parameter β. The removal of isolated images does help the performance, although is not as well as superimage replacement procedure.

tuning results. However, this would be too time-consuming to conduct. Experimental results manifest that our current setting also yield decent performance.

As is described in Sec. 3.2, either single or multiple holistic features can be fused with local features to enhance the inverted index. We consider three cases on UKbench dataset, *i.e*, employing semantic attribute only (SA), HSV color histogram only (HSV) and both of the two features (SA+HSV). Retrieval performances of the three different strategies are summarized in Fig. 6. It clearly proves that fusing multiple holistic features with our algorithm improves the retrieval precision dramatically. Quantitative study of the complementarity of holistic features and fuse them together guided by their complementarities is one of our future works.

We compare our method with recent retrieval approaches on UKbench and INRIA Holidays datasets, and summarize the results in Table 2. From the table, we can observe that the performance of our method is very competitive. The N-S score of UKbench increases 0.4 to 3.57 from 3.17, while on INRIA Holidays, mAP also has an improvement of 12.1 percent.

It is worthy to point out that we achieve the above performances in Table 2 without multiple feature extraction or re-ranking the results during online retrieval. Spatial verification or retrieval fusion, which introduce extra computations and memory costs, is adopted by some recent state-of-art retrieval systems, while our hybrid-indexing method deals with multi-type feature simultaneously totally during *off-line* stage. Our final performance is still decent considering we achieve them with a relatively low baseline. For example, [6] obtains a striking performance of N-S score on UKbench, but it is an increase of 0.23 based on a well implemented baseline which is 3.54.

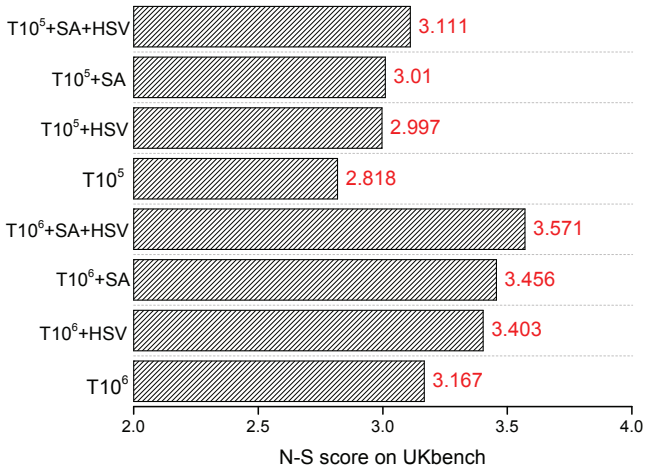

Fig. 6. Comparison of different fusion strategies. Complementarity of holistic features helps improvement of the retrieval precision on both baselines.

Another fact need to be stated is that compared with other indexing algorithms, memory consumption of hybrid-indexing is easily manageable. Indexed items of hybrid-indexing on UKbench counts to 7,869,590, while the conventional baseline indexes 5,712,698 items. That means an increase of 37.76% memory overhead emerges with a series of the algorithm procedures. After the superimage replacement and isolated image removal, the inverted index file size of UKbench rise from 47.3M to 66.0M. As for INRIA Holidays, indexed items show an increase of 24.13%, since less superimages are constructed on this dataset. Note that, some state-of-arts retrieval systems consume memory far more than ours, *e.g*, Hamming Embedding [19] storage an 64 bits binary signature along with the image id for each indexed item. Even omit to store TF (normally 32 bits) in its implementation, it is still a 100% memory consumption increase.

Some examples of retrieved images are presented in Fig. 7. It is obvious that our method is superior to retrieval algorithms that use single-type features.

Table 2. Comparison with the state-of-arts

Methods	Proposed	baseline($T10^6$)	[8]	[29]	[30]	[16]	[19]	[6]
Ukbench, N-S	3.57	3.17	3.60	3.56	3.52	3.45	3.42	3.77
Holidays,mAP(%)	78.6	66.5	80.9	78.1	76.2	N/A	81.3	84.6

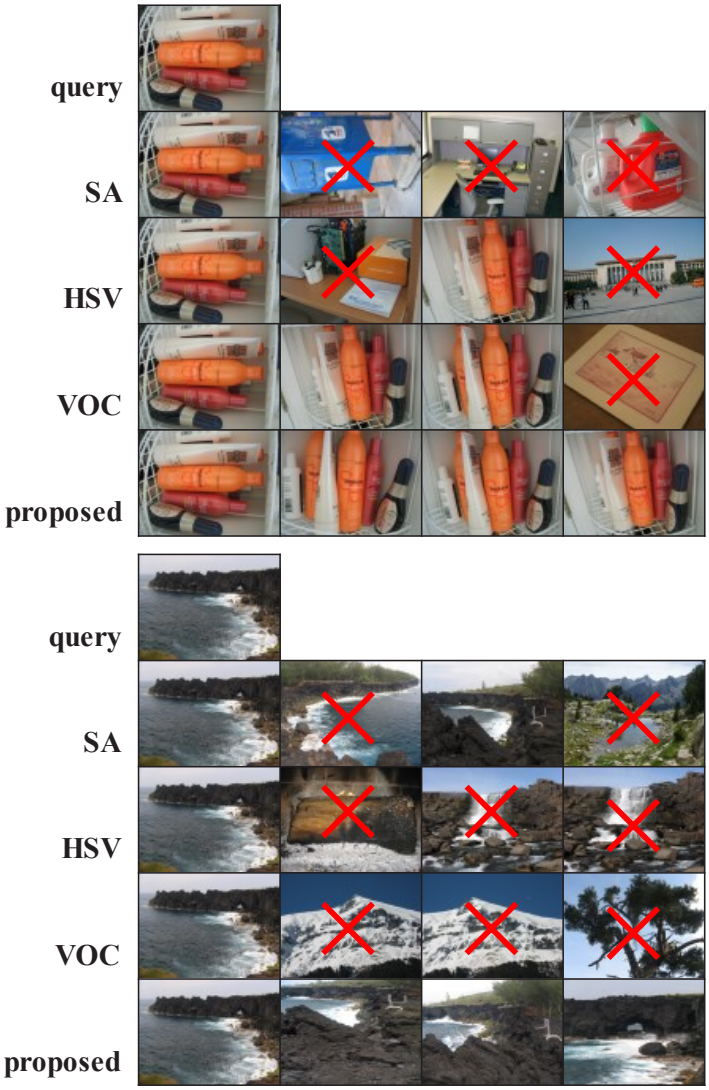

Fig. 7. Comparisons of retrieval results between our method and baseline methods, *i.e.*, image search with holistic feature semantic attribute (SA), color histogram (HSV) and BoWs retrieval baseline (VOC) on UKbench (*top*) and Holidays (*bottom*). The false results are marked by red crosses. The advantage of our method compared with any retrieval algorithms that use single-type features is evident.

4.3 Large-scale Image Search

To test the scalability of our approach, we also test our approach in the large-scale image search task. We employ a vocabulary tree with $B = 17$ and $T = 5$, which is trained on an independent large image dataset. We mix UKbench dataset with MIRFLICKR-1M [28] containing 1 million distraction images, and utilize images from UKbench as queries. Since our method is not restricted to the use of certain types of holistic features, we adopt Classemes [31] and HSV color histogram to fuse with local features to build hybrid-indexing. It is a more challenging task since superimages whose size larger than 1 on this dataset are rare due to its sparsity property. Despite of this, our method still improves the retrieval performance of baseline from 3.070 to 3.258 with a memory overhead of 19.7%. Consequently, we could conclude that our approach is also scalable to retrieve images from million-scale datasets.

5 Conclusion

In this paper, we present a novel *off-line* indexing approach to fuse multi-type cues including local and various holistic features simultaneously. By introducing complementary holistic cues into the classic inverted index, the proposed hybrid-indexing algorithm effectively combines two separated image search schemas, thus the retrieved results not only contain locally similar objects but also images that share relevant holistic characteristics. Experimental results manifest the promising advantages of the proposed method and warrant further investigation in this direction.

Acknowledgement. This work was supported in part to Dr. Qi Tian by ARO grant W911NF-12-1-0057 and Faculty Research Awards by NEC Laboratories of America. This work was supported in part by National Science Foundation of China (NSFC) 61429201.

References

1. Smeulders, A.W., Worring, M., Santini, S., Gupta, A., Jain, R.: Content-based image retrieval at the end of the early years. TPAMI **22** (2000) 1349–1380
2. Gehler, P., Sebastian, Nowozin: On feature combination for multiclass object classification. In: ICCV. (2009)
3. Zhang, S., Huang, J., Huang, Y., Yu, Y., Li, H., Metaxas, D.N.: Automatic image annotation using group sparsity. In: CVPR, IEEE (2010) 3312–3319
4. Fagin, R., Kumar, R., Sivakumar, D.: Efficient similarity search and classification via rank aggregation. In: ACM SIGMOD. (2003)
5. Jégou, H., Schmid, C., Harzallah, H., Verbeek, J.: Accurate image search using the contextual dissimilarity measure. TPAMI **32** (2010) 2–11
6. Zhang, S., Yang, M., Cour, T., Yu, K., Metaxas, D.N.: Query specific fusion for image retrieval. In: ECCV. Volume 2. (2012) 660–673
7. Ye, G., Liu, D., Jhuo, I.H., Chang, S.F.: Robust late fusion with rank minimization. In: CVPR. (2012)
8. Zhang, S., Yang, M., Wang, X., Lin, Y., Tian, Q.: Semantic-aware co-indexing for image retrieval. In: ICCV. (2013)

9. Luo, Q., Zhang, S., Huang, T., Gao, W., Tian, Q.: Superimage: Packing semantic-relevant images for indexing and retrieval. In: ICMR. (2014)
10. Zhou, W., Lu, Y., Li, H., Tian, Q.: Scalar quantization for large scale image search. In: Proceedings of the 20th ACM international conference on Multimedia, ACM (2012) 169–178
11. Lowe, D.G.: Distinctive image features from scale invariant keypoints. IJCV **60** (2004) 91–110
12. Rublee, E., Rabaud, V., Konolige, K., Bradski, G.: Orb: an effcient alternative to sift or surf. In: ICCV. (2011)
13. Zhang, S., Tian, Q., Huang, Q., Gao, W., Rui, Y.: Usb: Ultra short binary descriptor for fast visual matching and retrieval. Image Processing, IEEE Transactions on (2014)
14. Tian, Q., Sebe, N., Loupias, E., Huang, T., Lew, M.: Image retrieval using wavelet-based salient points. Journal of Electronic Imaging **10** (2001) 835–849
15. Nistér, D., Stewénius, H.: Scalable recognition with a vocabulary tree. In: CVPR. (2006)
16. Philbin, J., Chum, O., Isard, M., Sivic, J., Zisserman, A.: Object retrieval with large vocabularies and fast spatial matching. In: CVPR. (2007)
17. Zhang, S., Tian, Q., Huang, Q., Rui, Y.: Embedding multi-order spatial clues for scalable visual matching and retrieval. IEEE J. Emerg. Sel. Topics Circuits Syst. **4** (2014) 130–141
18. Chum, O., Philbin, J., Sivic, J., Isard, M., Zisserman, A.: Total recall: Automatic query expansion with a generative feature model for object retrieval. In: ICCV. (2007)
19. Jégou, H., Douze, M., Schmid, C.: Hamming embedding and weak geometric consistency for large scale image search. In: ECCV. (2008)
20. Zhang, S., Tian, Q., Hua, G., Huang, Q., Gao, W.: Descriptive visual words and visual phrases for image applications. In: ACM Multimedia. (2009)
21. Oliva, A., Torralba, A.: Modeling the shape of the scene: A holistic representation of the spatial envelope. IJCV **42** (2001) 145–175
22. Deng, J., Berg, A.C., Fei-Fei, L.: Hierarchical semantic indexing for large scale image retrieval. In: CVPR. (2011)
23. Farhadi, A., Endres, I., Hoiem, D., Forsyth, D.: Describing objects by their attributes. In: CVPR. (2009)
24. Douze, M., Ramisa, A., Schmid, C.: Combining attributes and fisher vectors for effcient image retrieval. In: CVPR. (2011)
25. Andoni, A., Indyk, P.: Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions. In: FOCS. (2006)
26. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: Imagenet large scale visual recognition challenge (2014)
27. Page, L., Brin, S., Motwani, R., Winograd, T.: The pagerank citation ranking: Bringing order to the web (1999)
28. Huiskes, M.J., Lew, M.S.: The mir flickr retrieval evaluation. In: MIR '08: Proceedings of the 2008 ACM ICMIR, New York, NY, USA, ACM (2008)
29. Wang, X., Yang, M., Cour, T., Zhu, S., Yu, K., Han, T.X.: Contextual weighting for vocabulary tree based image retrieval. In: ICCV. (2011)
30. Shen, X., Lin, Z., Brandt, J., Avidan, S., Wu, Y.: Object retrieval and localization with spatially-constrained similarity measure and k-NN reranking. In: CVPR. (2012)
31. Torresani, L., Szummer, M., Fitzgibbon, A.: Efficient object category recognition using classemes. In: ECCV. (2010) 776–789