Sparse Contextual Activation for Efficient Visual Re-Ranking

Song Bai, *Student Member, IEEE*, and Xiang Bai, *Senior Member, IEEE*

***Abstract*— In this paper, we propose an extremely efficient algorithm for visual re-ranking. By considering the original pairwise distance in the contextual space, we develop a feature vector called sparse contextual activation (SCA) that encodes the local distribution of an image. Hence, re-ranking task can be simply accomplished by vector comparison under the generalized Jaccard metric, which has its theoretical meaning in the fuzzy set theory. In order to improve the time efficiency of re-ranking procedure, inverted index is successfully introduced to speed up the computation of generalized Jaccard metric. As a result, the average time cost of re-ranking for a certain query can be controlled within 1 ms. Furthermore, inspired by query expansion, we also develop an additional method called local consistency enhancement on the proposed SCA to improve the retrieval performance in an unsupervised manner. On the other hand, the retrieval performance using a single feature may not be satisfactory enough, which inspires us to fuse multiple complementary features for accurate retrieval. Based on SCA, a robust feature fusion algorithm is exploited that also preserves the characteristic of high time efficiency. We assess our proposed method in various visual re-ranking tasks. Experimental results on Princeton shape benchmark (3D object), WM-SRHEC07 (3D competition), YAEL data set B (face), MPEG-7 data set (shape), and Ukbench data set (image) manifest the effectiveness and efficiency of SCA.**

***Index Terms*— Jaccard distance, feature fusion, re-ranking, retrieval, inverted index.**

I. Introduction

CONTEXTUAL similarity/dissimilarity [1]–[3] has been extensively exploited recently due to its effectiveness in various visual retrieval tasks, such as natural image search, shape retrieval, biological information retrieval, analysis of time series, *etc*.. Unlike traditional Content-based Image Retrieval (CBIR) systems that consider only pairwise dissimilarity measure for ranking and indexing, the approaches about contextual dissimilarity measure are proposed to explore the contextual information from the database instances, and enhance and refine the dissimilarity measure to improve the retrieval performance, which is usually considered as an unsupervised re-ranking procedure based on the given distance measure.

In general, the re-ranking procedure is often performed as a post-processing step of ranking initialization, which creates a ranking list for a given query. For image retrieval, the dissimilarity measure between a pair of images is often obtained by calculating the distance of their corresponding features under a certain metric. Given a query image, all the database images are sorted in an ascending/descending order according to their dissimilarities/similarities to the query. The ranking list for the query image can be finally initialized, where the most similar images occupy its top positions. A key issue in ranking initialization is to design proper features with enough discriminative power to represent an image, during which the Bag-of-Words (BoW) [4] image representation is often suggested.

Instead of ranking with pairwise dissimilarity measure, the contextual re-ranking algorithms have been proposed and demonstrate their effectiveness by considering the relationships among all database instances [1]–[3], [5]–[7]. In these re-ranking algorithms, the dissimilarity measure between two instances is iteratively updated and refined by taking into account their local distributions (neighborhood structure). As the neighborhood of each instance can be directly obtained from the ranking list, the key advantage of these re-ranking approaches is that training/labeled data is not required, operating in an unsupervised manner.

Though extensively studied, almost all the existing contextual re-ranking algorithms only pay much attention to the *effectiveness*, which refers to the level of retrieval accuracy. The *efficiency*, which refers to the time cost for the procedure of re-ranking, has been more or less neglected. However, both effectiveness and efficiency are quite important for a real-time retrieval system, and the tradeoff between them is badly required at present. The contextual re-ranking approaches [1], [8] often consider all the distances among instances of a given dataset, and the contextual dissimilarities are often achieved by operating on such a distance matrix. Therefore, a large computational effort is essential (typically, between $O(N^2)$ and $O(N^3)$, N is the number of images in the database), which seriously hinders their use in retrieval services that require for fast re-ranking. As a result, some re-ranking algorithms [9] are only applied to a subset of retrieved images to make a tradeoff between efficiency and effectiveness.

In this paper, we address the contextual re-ranking problem in an alternative and simpler manner. The contextual

Manuscript received March 13, 2015; revised June 13, 2015 and November 26, 2015; accepted December 22, 2015. Date of publication January 5, 2016; date of current version January 15, 2016. This work was supported in part by the National Natural Science Foundation of China under Grant 61573160 and Grant 61222308 and in part by the Program for New Century Excellent Talents in University under Grant NCET-12-0217. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Chunming Li. *(Corresponding author: Xiang Bai.)*

The authors are with the School of Electronic Information and Communications, Huazhong University of Science and Technology, Wuhan 430074, China (e-mail: songbai@hust.edu.cn; xbai@hust.edu.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2016.2514498

dissimilarity measure between two images is calculated by comparing two neighborhood sets, *i.e.* the neighborhood sets of a query and a target. It shares a similar intuition with the traditional approaches that the relationship between two images should not only be determined by the distance between them, but also influenced by their neighbors on the distance manifold. Our main contribution is to propose an extremely fast re-ranking algorithm called **Sparse Contextual Activation** (SCA) to compute the dissimilarity between such two neighborhood sets. Given a certain image, the basic idea of SCA is encoding its local distribution according to the original pairwise similarity into a single vector. Consequently, the contextual dissimilarity measure (set-to-set distance) between two images can be simply obtained by comparing two encoded vectors. Due to the sparsity property of such encoded vectors, the inverted index [10] can be further used to speed up the computation of the vector comparison for re-ranking. Besides the efficiency of SCA, it also achieves state-of-the-art retrieval performances on several standard benchmarks. In addition, we extend the proposed SCA for efficiently fusing multiple kinds of distance metrics for highly effective re-ranking while the inverted index can be also incorporated. Though the focus of this paper is re-ranking, contextual similarity has been successfully adopted in many image processing or vision tasks including image segmentation [11], [12], object tracking [13]. In this sense, the high efficiency of SCA is significant.

The rest of the paper is organized as follows. We first review some related work in Section II. The details of Sparse Contextual Activation (SCA) are introduced in Section III, and feature fusion based on SCA is described in Section IV. Experiments are carried out in Section V. Conclusions and future work are summarized in Section VI.

II. Related Work

Since there exist a large amount of works on re-ranking algorithms, we only review the unsupervised re-ranking algorithms in this section.

A. Re-Ranking With Single Distance Measure

In recent years, contextual information has been successfully explored to improve the retrieval accuracy by replacing a given pairwise similarity with a more faithful one, which is learned by considering the relationships among the database objects [3], [8], [14]–[18]. The contextual re-ranking has a very diverse taxonomy (graph transduction [3], [16], diffusion process [2], [8], [17], rank aggregation [19], [20], contextual similarity/dissmilaries measure [1], [14], [15], query expansion [21], [22], ranking list comparison [23], [24]). These post-processing approaches share the common spirit that the effectiveness of retrieval tasks is improved by taking use of relationships among dataset objects in an unsupervised manner, without labeled data.

One of the most classical algorithms is graph transduction (GT) [3]. As an unsupervised method, GT spreads the information from the labeled data to unlabeled data by regarding the query itself as the only labeled data.

A popular branch for re-ranking is diffusion process, which is summarized as a generic framework in [2]. Most variants of diffusion process share the same perspective that the pairwise similarity is context-sensitive, and the geometric structure of data manifold should be considered. In [8], Locally Constrained Diffusion Process (LCDP) is proposed to apply the affinity propagation with the constraint of locality. Tensor Product Graph (TPG) [17] diffuses the similarity information in the tensor product graph achieved by the tensor product of the original graph with itself.

Based on the observation that a good ranking is usually asymmetric, Contextual Dissimilarity Measure (CDM) [1] improves the retrieval performance of BoW vectors by modifying the neighborhood structure using Sinkhorn's scaling algorithm. Ranking consistency in [18] is used as a verification method to refine an existing ranking list. Query Expansion [21], [22] can substantially improve the retrieval performance by using relevant images as extra queries. Spatial verification [9] is proposed for re-ranking by considering the spatial constraints.

These aforementioned algorithms, as well as Self Diffusion (SD) [11], diffusion maps [25], aims at improving the retrieval accuracy, however the re-ranking efficiency is more or less neglected. For example, the time complexity of LCDP [8] and TPG [17] is $O(N^3)$. By contrast, our proposed SCA is a highly efficient re-ranking algorithm (in $O(N)$ time complexity), while also performs better in retrieval accuracy.

B. Re-Ranking With Multiple Distance Measures

Multiple distance measure fusion has been proven effective in various visual applications, such as visual tracking [13], image classification [26] *etc*.. Considering that one distance measure only focuses one aspect of images, some re-ranking algorithm also deals with multiple complementary features.

Zhang *et al*. [27], [28] fuse BoW feature and holistic feature by a graph-based query specific fusion, and re-ranking is performed by using the local PageRank algorithm or finding the weighted maximum density subgraph. Co-transduction [29] adopts a semi-supervised framework based on co-training [30], [31] to combine complementary features for image and shape retrieval. In [32], a regularization-based feature selection algorithm is proposed to leverage both the sparsity and clustering properties of multiple features.

Besides multiple distance measure fusion, some algorithms also focus on designing an individual distance measure using multiple features. For example, c-MI [33] integrates SIFT descriptor [34] and color descriptor into a multi-dimensional inverted index. In [35], a multi-IDF scheme is proposed, by which different binary features are coupled into the inverted index. Co-indexing [36] jointly embeds local invariant features and semantic attributes.

Based on sparse contextual activation, we also propose a re-ranking version that deals with multiple distance measures, which not only maintains the characteristic of efficiency, but also improves the performance significantly.

III. Sparse Contextual Activation

Let $X = \{x_1, x_2, \ldots, x_N\}$ denote a collection of images. We define two functions listed as follows:

- Function $f : x \to \mathbb{R}^n$: it extracts a n-dimensional feature to represent the input image x.
- Function $d : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$: it computes the distance for the two input feature vectors $f(x_q)$ and $f(x_p)$ under a certain metric, where x_q and x_p represent two images in the database.

The distance $d\left(f(x_q), f(x_p)\right)$ is taken as the dissimilarity of the images x_q and x_p. To simplify the notation, we use $d(x_q, x_p)$ to replace $d\left(f(x_q), f(x_p)\right)$ below where possible.

After all the pairwise dissimilarity values related to the given query x_q are achieved, we could initialize the ranking list by sorting the dissimilarity in an increasing order. The images with smaller dissimilarity values are ranked higher in the retrieval list, and vice versa.

Next, we introduce our proposed re-ranking algorithm to refine the original distance measure with high time efficiency.

A. Preliminary

We deem that if x_q and x_p are similar, their retrieval results, especially the top-ranked results, are exactly or approximately the same in the expected situation. Let $\mathcal{N}_k(x_q)$ represent the *neighborhood set* of x_q achieved by the k-nearest neighbors rule in the original distance space d. $\mathcal{N}_k(x_q)$ is a mathematical set that contains the top-k candidates in the ranking list of x_q. Then, the distance of two neighborhood sets $\mathcal{N}_k(x_q)$ and $\mathcal{N}_k(x_p)$ is measured by Jaccard distance as

$$d_J(x_q, x_p) = 1 - \frac{|\mathcal{N}_k(x_q) \cap \mathcal{N}_k(x_p)|}{|\mathcal{N}_k(x_q) \cup \mathcal{N}_k(x_p)|}, \tag{1}$$

where $|.|$ calculates the cardinality of the input set, and $|\mathcal{N}_k(x_q)| = k$. With the distance measure defined by Eq. (1), the original ranking list of a given query x_q can be re-ranked.

The re-ranking distance measure in Eq. 1 is expected to achieve better performance than the original one, for it utilizes the additional contextual information as diffusion process does. However it also has many shortcomings.

1) The neighbors in the neighborhood set contribute equally. It is not a proper behavior, since the top-ranked neighbors are more likely to be true positive patterns. Assigning larger weights to the top-ranked neighbors, and increasing their effects on the re-ranking distance measure is more reasonable.
2) The re-ranking distance measure is defined between two sets. In the specific scenario of re-ranking, it is more convenient to define the distance measure on two vectorial features.
3) The neighborhood set is simply defined as the k-nearest neighbors, which cannot guarantee that the images from the same category could own similar neighborhood sets, especially when a certain amount of outliers also occupy top positions in the ranking list.

In the next section, Sparse Contextual Activation (SCA) is proposed to address these problems. The Jaccard distance defined in Eq. (1) will serve as the baseline method, and we will compare it with our proposed SCA in terms of retrieval performance and running time. For notation clarity, we refer to the baseline method as Jaccard re-ranking below.

B. The Proposed Sparse Contextual Activation

The neighborhood set $\mathcal{N}_k(x_q)$ is converted to a vector representation by defining an binary *indicator function* $F_q = [F_{q,1}, F_{q,2}, \ldots, F_{q,N}]$ as

$$F_{q,p} = \begin{cases} 1 & \text{if } x_p \in \mathcal{N}_k(x_q) \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

As we can see, the binary vector F_q shows whether a certain image x_p appears in the neighborhood set of x_q.

Based on the definition of the indicator function, the intersection and union set of $\mathcal{N}_k(x_q)$ and $\mathcal{N}_k(x_p)$ are interpreted as

$$\mathcal{N}_k(x_q) \cap \mathcal{N}_k(x_p) \Longleftrightarrow MIN\left(F_q, F_p\right), \tag{3}$$

$$\mathcal{N}_k(x_q) \cup \mathcal{N}_k(x_p) \Longleftrightarrow MAX\left(F_q, F_p\right), \tag{4}$$

where MIN (or MAX) calculates the element-wise minimum (or maximum) value for two input vectors of the same length. Thus we attain the cardinality for the intersection and the union set by computing the L_1 norm of their corresponding indicators as

$$|\mathcal{N}_k(x_q) \cap \mathcal{N}_k(x_p)| = \|MIN(F_q, F_p)\|_1, \tag{5}$$

$$|\mathcal{N}_k(x_q) \cup \mathcal{N}_k(x_p)| = \|MAX(F_q, F_p)\|_1. \tag{6}$$

Based on Eq. (5) and Eq. (6), we can rewrite the definition of Jaccard distance in Eq. (1) as

$$\hat{d}_J(x_q, x_p) = 1 - \frac{\sum_{i=1}^{N} \min\left(F_{q,i}, F_{p,i}\right)}{\sum_{i=1}^{N} \max\left(F_{q,i}, F_{p,i}\right)}. \tag{7}$$

Now, the definition of Jaccard distance in Eq. (1) has been successfully defined in vector space. That is to say, we do not use a set, but use an indicator function to represent the neighbors of an image. As a result, the Jaccard distance of two neighborhood sets can be easily achieved by vector comparison through Eq. (7).

Note that the indicator function in Eq. (2) also considers the neighbors equally, but it is easy to implement different weights by restricting the value of the original binary indicator vector in the unit interval [0, 1]. However, it may be misleading that $F_{q,p}$ is assigned to a certain constant between 0 and 1, since the role of $F_{q,p}$ is to indicate the membership of the image x_p in the neighborhood set $\mathcal{N}_k(x_q)$. In classical set theory, the membership of x_p in the set $\mathcal{N}_k(x_q)$ is exact (x_q either belongs or does not belong to the set). It seems difficult to generalize the binary indicator function in the unit interval $[0, 1]$ with rational explanations.

To tackle with the problem, we introduce the *fuzzy set theory*. In mathematics, fuzzy set is a set whose elements have degrees of membership determined by a *membership function*. Compared with the classical set theory, fuzzy set allows the gradual assessment of the membership of elements in a set. At last, we define the neighborhood set as a fuzzy set in fuzzy set theory, and the membership grade of x_p in the

neighborhood set $\mathcal{N}_k(x_q)$ is determined by the corresponding membership function $F_{q,p} \in [0, 1]$.

The problem we face now is how to determine the membership grade of neighbors. A natural solution is to use the elements in $\mathcal{N}_k(x_q)$ to reconstruct x_q in the feature space with non-negative constraint, and the weights for reconstruction can be used as the membership grades. It can be formulated as

$$\min \sum_q \| f(x_q) - \sum_{i|x_i \in \mathcal{N}_k(x_q)} F_{q,i} f(x_i) \|^2,$$
$$s.t.\ 1^{\mathrm{T}} F_q = 1, F_q \succeq 0. \tag{8}$$

Except for the non-negative constraint, this formulation almost shares the same perspective with Locally Linear Embedding (LLE) [37], a classical non-linear dimension reduction algorithm. LLE expects each data point and its neighbors lie on or close to a locally linear patch of the manifold, and the reconstruction weights are used for dimension reduction by a neighborhood preserving mapping.

However, in the specific scenario, the above solution may be not fit enough for visual re-ranking for three reasons. First, LLE usually presumes that there is sufficient data so that the data manifold is well-sampled, but retrieval task may also be needed in small datasets. Second, the requirement for real-time retrieval is usually declared, and it is time-consuming to solve the least square optimization problem for each query presented in Eq. (8). On the other hand, the image is represented by a set of vectors instead of a single vector in some cases (*e.g.* shape analysis in [38]–[40]), which makes the above optimization problem more difficult to solve. Hence, we just apply the (truncated) Gaussian kernel to the pairwise distances with the given query x_q, and the membership function is defined as

$$F_{q,p} = \begin{cases} \exp\left(-d(x_q, x_p)\right) & \text{if } x_p \in \mathcal{N}_k(x_q) \\ 0 & \text{otherwise.} \end{cases} \tag{9}$$

As a result, the top-ranked neighbors are assigned larger membership grades. F_q is subsequently L_1 normalized according to the sum-to-one constraint, and used for re-ranking through Jaccard distance defined in Eq. (7). In this way, the Jaccard distance is generalized to non-binary vectors.

In summary, the neighbors of x_q actually act as *Local Coordinate System*, and we can get a *Sparse Contextual Activation* denoted by F_q for x_q through Eq. (9). The "Contextual" here indicates that F_q has non-zero values only in the index where the neighbors of x_q are located. Usually the cardinality of $\mathcal{N}_k(x_q)$ is much smaller than the size of the entire dataset, so the contextual activation is also a sparse vector. The property of sparsity is crucial in our algorithm. With this constraint, the negative influences of unreliable references are eliminated. What is more important is that the calculation of Jaccard distance in Eq. (7) can be accelerated using inverted file as presented in Section III-C. In Fig. 1, we give an illustration of our proposed Sparse Contextual Activation (SCA).

C. Inverted Index Embedding

The proposed Sparse Contextual Activation (SCA) already manages assigning different weights to the neighbors at different positions in the ranking list, and gives a vector representation used for re-ranking. However, distance computation between a pair of SCAs is also waste of time, especially when the size of database becomes larger. Although the length of SCA is equal to the size of image database N, but the number of non-negative values in SCA is independent, only determined by the cardinality of neighborhood set. Considering the sparsity property of SCA, we introduce the inverted index [10] to reduce the computation complexity significantly.

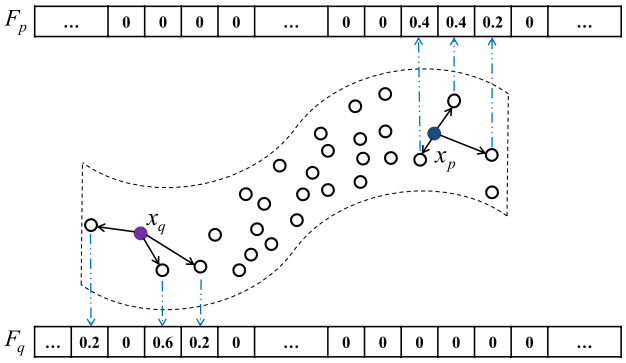


Fig. 1. The illustration of sparse contextual activation.

Inverted index is a scalable indexing structure to store a large collection of images with their features. Although it has been applied to image retrieval successfully (*e.g.* [41]–[43]), it is the first work that introduces inverted index to visual re-ranking to our best knowledge now. Moreover, different from the usage of inverted index in Minkowski metric (Euclidean distance, Manhattan distance, *etc.*), we prove the feasibility of applying it in the metric of Jaccard distance theoretically.

Given two sparse contextual activations F_q and F_p, the MIN operation can be computed as

$$\begin{aligned} \|MIN(F_q, F_p)\|_1 = & \sum_{i|F_{q,i}\neq 0, F_{p,i}\neq 0} \min(F_{q,i}, F_{p,i}) \\ & + \sum_{i|F_{q,i}=0} \min(F_{q,i}, F_{p,i}) \\ & + \sum_{i|F_{p,i}=0} \min(F_{q,i}, F_{p,i}). \end{aligned} \tag{10}$$

Since the sparse contextual activation only contains non-negative values, it is easy to find that last two items in Eq. (10) are equal to zero. So the MIN operation can be achieved much more efficiently by

$$\|MIN(F_q, F_p)\|_1 = \sum_{i|F_{q,i}\neq 0, F_{p,i}\neq 0} \min(F_{q,i}, F_{p,i}), \tag{11}$$

and it is only a query-dependent operation.

By contrast, the computation of the MAX operation seems to be a bit complicated, since the item $\sum_{i|F_{q,i}=0} \max(F_{q,i}, F_{p,i}) = \sum_{i|F_{q,i}=0} F_{p,i}$ is not only determined by the query side. However, we also offer an efficient way to calculate the MAX operation. Note that

$$\|F_q\|_1 + \|F_p\|_1 = \|MIN(F_q, F_p)\|_1 + \|MAX(F_q, F_p)\|_1.$$

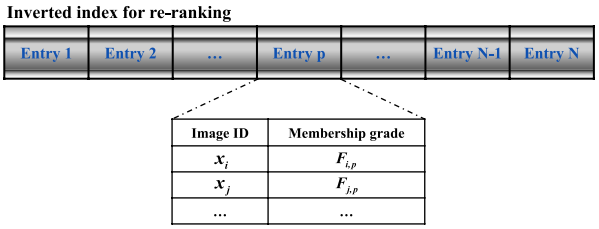


Fig. 2. The inverted index for re-ranking.

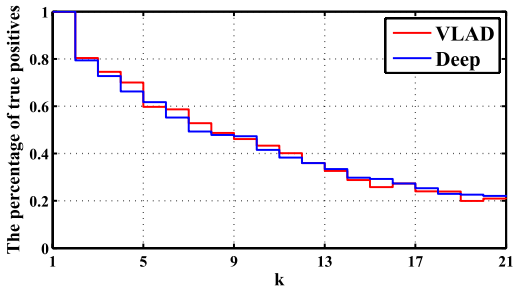


Fig. 3. The distribution of the percentage of true positives at the k-th position in the ranking list for all the queries. "VLAD" and "Deep" denote two visual features used in our experiment. Please refer to Section V-A for more details.

For two L_1 normalized sparse contextual activation, we can get

$$\begin{aligned}\|MAX(F_q, F_p)\|_1 &= 2 - \|MIN(F_q, F_p)\|_1, \\ &= 2 - \sum_{i|F_{q,i}\neq 0, F_{p,i}\neq 0} \min(F_{q,i}, F_{p,i}). \quad (12)\end{aligned}$$

In summary, our structure of inverted index is built as follows: (1) It has N entries as Fig. 2 shows, where N is the size of database. Each entry relates to an image that acts as a base for activation. (2) For each entry p, we store the IDs of images whose neighborhood sets contain x_p and the corresponding membership grades. In other words, x_p owns non-zero membership grades in these neighborhood sets. (3) When re-ranking for the query x_q, distance computation can be conducted in a much smaller space using Eq. (11) and Eq. (12) as inverted index usually does.

Different from the inverted index on local descriptors in the paradigm of Bag of Words (BoW) model [4], our inverted index for re-ranking can be assumed as the second-level index, which is used in the level of images.

D. Local Consistency Enhancement

The images from the same category are expected to own the same neighborhood set, so that the distances between their sparse contextual activations are small. However, the neighborhood set is determined using k-nearest neighbors rule, which may be the simplest and the most efficient principle. Indeed, one can identify the neighborhood set by using more sophisticated rules (*e.g.* dominant neighbors [17]), but it will increase the computational time dramatically. Any extra time cost is not what we want in the proposed algorithm.

Inspired by the query expansion [21], [22] and local relevance feedback [44] in information retrieval, we propose to enhance the local consistency in generating SCAs of images from the same category using a similar way. In more detail,

Algorithm 1 The Pseudocode of SCA

Input:
$D = \{d(x_q, x_p)\}, \forall x \in X$: the original distance matrix;
k_1: the size of neighborhood set;
k_2: the parameter in local consistency enhancement.
Output:
$\hat{D}_J = \{\hat{d}_J(x_q, x_p)\}, \forall x \in X$: the refined distance matrix;
begin
  for $i = 1$ *to* N **do**
    Determine $\mathcal{N}_k(x_i)$ using k_1.
    Compute F_i according to Eq. (9).
  if $k_2 \neq 1$ **then**
    for $i = 1$ *to* N **do**
      Determine $\mathcal{N}_k(x_i)$ using k_2.
      Enhance F_i according to Eq. (13).
  Build inverted file on $F_i, 1 \leq i \leq N$;
  for $x_q \in X$ **do**
    Determine non-zero entries of x_q in inverted file.
    Estimate $\hat{d}_J(x_q, x_p)$ according to Eq. (11) and Eq. (12), $\forall x_p \in X$.
  return $\hat{D}_J$

we define the Local Consistency Enhancement (LCE) on the sparse contextual activation as

$$F_q := \frac{1}{|\mathcal{N}_k(x_q)|} \sum_{i \in \mathcal{N}_k(x_q)} F_i, \quad (13)$$

by speculating that the images which are located very high in the initial ranking list of the query x_q are from the same category as the query. In our experiment, LCE is applied only once, both in the query side and database side.

In order to confirm our conjecture, an empirical analysis is performed on Princeton Shape Benchmark (PSB) [45]. We plot the percentage of true positives at the k-th position in the ranking list for all the queries in Fig. 3. It can be seen that the percentage of true positives decreases with the value of k increasing, and it is extremely high when k is small. Note that several false positives also exist at smaller value of k, which is known as *query drift*, but the small percentage of false positives will not impair the whole performance too much.

Based on the above analysis, the proposed LCE is an unsupervised method that does not need to know the labels. Compared with sparse contextual activation that considers the information from the direct neighbors of x_q, LCE is defined to consider the neighbors of the second order (*i.e.* the neighbors of the neighbors of x_q). As a result, LCE is more likely to include noise and outliers. Hence, we restrict the size of the neighborhood set used in LCE to a much smaller value in the experiments.

In order to distinguish the two neighborhood sets, both derived from the original distance space d, used in sparse contextual activation (Eq. (9)) and local consistency enhancement (Eq. (13)), we denote the size of the former one as k_1 and that of the latter one as k_2 below.

The whole algorithm is summarized in Algorithm 1.

IV. Sparse Contextual Activation With Multiple Distance Measures

The proposed Sparse Contextual Activation presented above deals with only one input distance measure. In this section, we present how to introduce our method into rank aggregation for multiple input distance measures. We take two input distance measures as an example.

Let f^α and f^β denote two feature extractor functions for images, and d^α and d^β represent the corresponding distance functions respectively. Then for the image x_q with the feature extractor f^α (or f^β), we can also achieve the neighborhood set $\mathcal{N}_k^\alpha(x_q)$ (or $\mathcal{N}_k^\beta(x_q)$).

A. High Set and Low Set

Given a query image x_q and its two neighborhood sets $\mathcal{N}_k^\alpha(x_q)$ and $\mathcal{N}_k^\beta(x_q)$. We call the intersection set of the two neighborhood sets *High Set*, which can be described as

$$\mathcal{N}_k^{(H)}(x_q) = \mathcal{N}_k^\alpha(x_q) \cap \mathcal{N}_k^\beta(x_q), \tag{14}$$

and call the union set *Low Set* as

$$\mathcal{N}_k^{(L)}(x_q) = \mathcal{N}_k^\alpha(x_q) \cup \mathcal{N}_k^\beta(x_q). \tag{15}$$

The "high" indicates that the images in the high set are more likely to be true positives, since the images in the set occupy top positions in two ranking lists achieved by two distance functions d^α and d^β. By contrast, the low set may contains more false positives, for it is defined with a looser constraint, *i.e.* the images will be assigned to the low set as long as they are ranked high by either d^α or d^β.

In order to perform extremely fast rank aggregation later, we also generate the sparse contextual activations for high set and low set similarly. In more detail, let F_q^α (or F_q^β) denote the sparse contextual activation of x_q computed using Eq. (9) under the distance measure d^α (or d^β). The membership functions of high set and low set are defined respectively based on Eq. (3) and Eq. (4)

$$\mathcal{N}_k^{(H)}(x_q) \Longleftrightarrow F_q^{(H)} = MIN\left(F_q^\alpha, F_q^\beta\right), \tag{16}$$

$$\mathcal{N}_k^{(L)}(x_q) \Longleftrightarrow F_q^{(L)} = MAX\left(F_q^\alpha, F_q^\beta\right). \tag{17}$$

As a result, two sparse contextual activations $F_q^{(H)}$ and $F_q^{(L)}$, corresponding with high set and low set respectively, are achieved for x_q.

One can find that we perform feature fusion naturally by defining the two sets, so the unsupervised estimation about the weights of two individual features during fusion procedure are avoided. As we all know, different features are often in different scales or measured by different statistics, which makes the weight learning in feature fusion difficult, especially in an unsupervised way. Our proposed sparse contextual activation for feature fusion is simple yet delicate. It does not require any complicated learning or optimization methods.

B. Distance Fusion

After getting two sparse contextual activations for each image in the database, we define the distance measure for rank aggregation as

$$\hat{d}_J(x_q, x_p) = 1 - \frac{1}{2}\sum_{I=H,L} \frac{\sum_{i=1}^N \min\left(F_{q,i}^{(I)}, F_{p,i}^{(I)}\right)}{\sum_{i=1}^N \max\left(F_{q,i}^{(I)}, F_{p,i}^{(I)}\right)}. \tag{18}$$

The contribution of high set and low set are treated equally, which avoids the parameter tuning at the stage. It seems that the individual performance of high set is better than low set, due to its high percentage of true positives. The true fact is that the definition of high set may be too strict in the specific scenario of re-ranking. In some cases (especially when parameter k_1 is not large enough), a much small number of images will be assigned to the high set, which deprives its discriminative ability. However, with k_1 increasing, the individual performance of high set will surpass low set after a certain threshold without doubt. We will experimentally analyse the performance difference between high set and low set in Section V-E, and prove that a simple linear combination of them is better than using either one only.

It should be mentioned that the inverted index (Section III-C) and local consistency enhancement (Section III-D) can also be introduced into this feature fusion paradigm. The two additional methods can improve the efficiency and accuracy remarkably.

V. Experiments

In this section, we will evaluate the performance of the proposed Sparse Contextual Activation (SCA) in various visual re-ranking task. The datasets we use are Princeton Shape Benchmark (PSB) [45], Watertight Models track of SHape REtrieval Contest 2007 dataset (WM-SHREC07) [46], YALE face dataset B [47], and MPEG-7 shape dataset [48] and Ukbench image dataset [10]. The discussion about the parameter settings and the analysis of the algorithm complexity are presented in Section V-E and Section V-F respectively.

A. 3D Object Retrieval

In this section, we show the application of the proposed SCA to 3D object retrieval on the well-known Princeton Shape Benchmark (PSB) [45] and Watertight Models track of SHape REtrieval Contest 2007 dataset (WM-SHREC07) [46].

The PSB benchmark contains 1,804 3D polygonal models, which are divided into training set and testing set with 907 models each. Following the common settings, only the testing set is used to evaluate the performance of 3D object retrieval. The testing set is spilt into 92 categories, and the number of models per category ranges from 4 to 50.

SHape REtrieval Contest (SHREC) is the most authoritative competition for evaluating the effectiveness of 3D object retrieval algorithms. It will be held each year, involving multiple tracks, such as sketch-based 3D retrieval, textured 3D retrieval, *etc*.. In this paper, WM-SHREC07 is chosen, which consists of 400 watertight mesh models that are evenly distributed into 20 classes. The models exhibit sufficient and

TABLE I

THE PERFORMANCE COMPARISON IN FT OF DIFFERENT RE-RANKING ALGORITHMS ON PSB DATASET AND WM-SHREC07 DATASET

	PSB dataset				WM-SHREC07 competition			
Methods	VLAD	Gain	Deep feature	Gain	VLAD	Gain	Deep feature	Gain
Baseline	0.575	-	0.572	-	0.708	-	0.783	-
SD [11]	0.576	+0.17%	0.582	+1.75%	0.711	+0.42%	0.760	-2.94%
GT [3]	0.583	+1.39%	0.600	+4.90%	0.766	+8.19%	0.769	-1.78%
TPG [17]	0.613	+6.61%	0.598	+4.55%	0.761	+7.49%	0.839	+7.15%
LCDP [8]	0.617	+7.30%	0.603	+5.42%	0.766	+8.19%	0.843	+7.66%
SCA	**0.660**	**+14.78%**	**0.617**	**+7.87%**	**0.795**	**+12.23%**	**0.876**	**+11.88%**

TABLE II

THE PERFORMANCE COMPARISON WITH OTHER STATE-OF-THE-ART ALGORITHMS ON PSB DATASET AND WM-SHREC07 DATASET

	PSB dataset				WM-SHREC07 competition			
Methods	NN	FT	ST	DCG	NN	FT	ST	DCG
LFD [52]	0.657	0.380	0.487	0.643	0.923	0.526	0.662	-
Tabia *et al*. [53]	-	-	-	-	0.853	0.527	0.639	0.719
DESIRE [54]	0.665	0.403	0.512	0.663	0.917	0.535	0.673	-
tBD [55]	0.723	-	-	0.667	-	-	-	-
Covariance [56]	-	-	-	-	0.930	0.623	0.737	0.864
2D/3D Hybrid [57]	0.742	0.473	0.606	-	0.955	0.642	0.773	-
PANORAMA [44]	0.753	0.479	0.603	-	0.957	0.673	0.784	-
Shape Vocabulary [58]	0.717	0.484	0.609	0.723	-	-	-	-
PANORAMA + LRF [44]	0.752	0.531	0.659	-	0.957	0.743	0.839	-
3DVFF [59]	-	-	-	0.841	-	-	-	-
SCA+Deep	0.773	0.617	0.740	0.800	0.952	0.876	0.951	0.957
SCA+VLAD	0.803	0.660	0.786	0.826	0.955	0.795	0.911	0.938
SCA+VLAD+Deep	**0.837**	**0.700**	**0.810**	**0.850**	**0.990**	**0.900**	**0.956**	**0.972**

diverse variation, from pose change to shape variability in the same semantic category.

Four evaluation metrics are adopted to assess the retrieval performance, listed as follows:

- Nearest Neighbor (NN): the percentage of the closest matches that belongs to the same class as the query.
- First Tier (FT): the recall for the top $C-1$ matches in the ranked list, where C is the number of shapes in the category that query belongs to.
- Second Tier (ST): the recall for the top $2(C-1)$ matches in the ranked list, where C is the number of shapes in the category that query belongs to.
- Discounted Cumulative Gain (DCG): a statistic that attaches more importance to the correct results near the front of the ranked list than the correct results at the end of the ranked list, under the assumption that a user is more likely to consider the retrieved candidates in the front of the list.

Please refer to [45] for more details about the definition of NN, FT, ST and DCG if needed. The values of all the aforementioned metrics range from 0 to 1, and larger values indicate better performance.

In the 3D object retrieval task, we use two view-based baseline methods (one is Vector of Aggregated Local Descriptor (VLAD) [49], and the other one is deep feature learned with Convolutional Neural Network (CNN)). For VLAD, we use the same pipeline as [50]. The number of depth views is set to 64, and the codebook size is 2048. For the deep feature, we utilize the descriptor proposed in [51].

We first demonstrate the performance of different re-ranking algorithms using the same baseline methods (*i.e.*, VLAD and deep feature) in Table I, and FT is chosen as the evaluation metric. When computing SCA, we fix the size of neighborhood set k_1 to 10 for PSB dataset and 17 for WM-SHREC07 dataset, and the parameter k_2 for local enhancement to 4. We report the performances of some typical re-ranking algorithms (Self diffusion (SD) [11], Tensor Product Graph (TPG) [17], Locally Constrained Diffusion Process (LCDP) [8]) in the optimal parameter setup.

It can be observed that our proposed SCA achieves the best performance among all the compared re-ranking methods. SD obtains the worst results since it loses the constraint of "locality" used in LCDP and TPG. The percent gain in performance of LCDP and TPG in PSB dataset is not as large as in WM-SHREC07. It can be explained in two aspects: the first one is that the baseline in PSB dataset is much lower than that in WM-SHREC07 dataset, which means that it will include more noise and outliers in the neighborhood set; the second one is that the number of objects per category in PSB dataset is various. By contrast, the distribution of objects in WM-SHREC07 is exactly balanced, *i.e.* 20 objects are assigned to one category. The imbalanced distribution of objects in PSB dataset makes kNN rule difficult to generalize well. Nevertheless, our proposed SCA also performs stably and well in both datasets although kNN is used to define the neighborhood set.

The comparison with other state-of-the-art algorithms are presented in Table II. As we can see, our proposed

SCA achieves the new state-of-the art performance among all other algorithms for all the four evaluation metrics by fusion VLAD and deep feature in both datasets. PANORAMA is one of the most representative 3D shape descriptors in recent years, and our proposed SCA outperforms it by 8.4% in NN, 22.1% in FT and 20.7% in ST in PSB dataset. In [44], the performance of PANORAMA is improved by large margins using Local Relevance Feedback (LRF). LRF also exploits the contextual contribution as SCA, and the superior results reveal the effectiveness of our proposed method.

Among the all compared methods, 2D/3D Hybrid [57], PANORAMA [44] and 3DVFF [59] consider the fusion of multiple complementary features in the hope of more robust distance measure. 2D/3D Hybrid just simply concatenates 2D features based on depth buffers and 3D features based on spherical harmonics, and the dissatisfactory performance verifies the importance of designing more discriminative feature fusion methods. 3DVFF employs Multi-Feature Anchor Manifold that approximates multiple manifolds of heterogeneous features, which can be assumed as one variant of diffusion-based re-ranking methods. SCA also leads to a better retrieval accuracy than 3DVF in DCG (DCG is the only widely-accepted evaluation metric adopted in [59]).

B. Face Retrieval

We also evaluate the performance of SCA in face retrieval on YALE face dataset B [47]. YALE face dataset B is a standard benchmark widely used for face clustering, which is composed of face images sharing various poses and illumination conditions. In order to keep the comparison fair, we use the same subset and the same baseline method as generic diffusion framework [2]. Specifically, 15 subjects with 11 different conditions are gathered to generate a new dataset. Each image is normalized to 0-mean and 1-variance, and Euclidean distance between the vectorized representations is adopted to measure the pairwise dissimilarity directly. The evaluation metric is bull's eye score, which counts the recall before top-15 ranking list.

The baseline bull's eye score for the selected subset is 69.48%. Note that our goal is not achieving superior retrieval performance in this dataset, since better performance can be obtained easily by using more discriminative descriptors, such as LBP [60], *etc*.. Instead, we focus on the performance improvement by our proposed re-ranking method. In [2], a classical branch of re-ranking methods called diffusion process is summarized, so the comparison with these methods will be more convincing. Similar to SCA, diffusion-based re-ranking methods capture the structure of the underlying data manifold by using the contextual information. The primary difference is that they propagate the affinity values with random walks on a pre-defined graph in an iterative manner, while SCA does not need the iterative procedure that is often of great computational cost. We refer the readers to [2] for more details about diffusion process if necessary.

In Table III, the retrieval performances resulting from other state-of-the-art methods and our proposed method are reported. All the numerical results are borrowed from [2] or computed by using the public-available codes. When computing our sparse contextual activation, we manually set the size of neighborhood set k_1 to 4, and the parameter k_2 for local consistency enhancement to 5.

TABLE III
THE PERFORMANCE COMPARISON WITH OTHER STATE-OF-THE-ART ALGORITHMS ON THE YALE FACE DATASET B

Algorithm	Bull's eye score
Graph Transduction [3]	68.70%
Self Diffusion [11]	71.46%
Tensor Product Graph [17]	75.32%
Locally Constrained Diffusion Process [8]	75.59%
Generic diffusion [2]	77.30%
SCA	**77.80%**

TABLE IV
THE PERFORMANCE COMPARISON WITH OTHER STATE-OF-THE-ART ALGORITHMS ON MPEG-7 DATASET

Descriptor	Re-ranking algorithm	Score
SC	-	86.80%
	GT [3]	92.91%
	SCA	**95.21%**
IDSC	-	85.40%
	CDM [1]	88.30%
	Indexing Re-Ranking [6]	91.56%
	GT [3]	91.61%
	Pairwise Recom. [61]	92.21%
	RL-Sim [19]	92.62%
	LCDP [8]	93.32%
	SSP [5]	93.35%
	SCA	**93.44%**
SC+IDSC	-	92.16%
	Co-T [29]	97.72%
	LCMD [7]	98.84%
	SCA	**99.01%**

As we can see from the table, SCA outperforms Self Diffusion (SD) [11], Tensor Product Graph (TPG) [17], Locally Constrained Diffusion Process (LCDP) [8] by 6.34%, 2.48% and 2.21%. As for the generic diffusion process, it enumerates all the variants of diffusion process by using four different types of initialization, six different types of transition and matrices and three different update schemes. The bull's eye score of the generic version of diffusion process is chosen as the best performance for all the variants. However, SCA also achieves 77.80%, which is the best performance among all re-ranking methods, including the generic diffusion.

C. Shape Retrieval

Finally, we test our method for shape retrieval on MPEG-7 dataset [48]. It consists of 1,400 binary images divided into 70 categories evenly. Bull's eye score is used to evaluate the performance, which counts the recall in the top-40 ranking list. We use the same baseline methods as Co-transduction [29], where Shape Context (SC) [38] and Inner Distance Shape Context (IDSC) [39] are used as the raw descriptors.

The comparison with other state-of-the-art algorithms is presented in Table IV. In order to improve the readability of the comparison, the table is divided into three parts with

TABLE V

THE PERFORMANCE COMPARISON WITH STATE-OF-THE-ART METHODS IN UKBENCH DATASET. THE DIFFERENCE BETWEEN [33]^a, [35]^a AND [33]^b, [35]^b IS THE METHODS WITH SUPERSCRIPT "b" USE THE ADDITIONAL GRAPH FUSION ALGORITHM PROPOSED IN [27]. [27]^a AND [27]^b DENOTE GRAPH PAGERANK AND GRAPH DENSITY RESPECTIVELY

Algorithms (single-feature)					
[43]	[49]	[63]	[64]	[17]	[41]
3.45	3.47	3.53	3.56	3.61	3.64
[65]	[1]	SCA+HSV		SCA+SIFT	
3.67	3.68	3.56		**3.69**	
Algorithms (multi-feature)					
[36]	[35]^a	[66]	[29]	[67]	[33]^a
3.60	3.60	3.62	3.66	3.68	3.71
[27]^a	[27]^b	[35]^b	[33]^b	SCA+SIFT+HSV	
3.76	3.77	3.79	3.85	**3.86**	

each part using the same baseline method. The baseline of the direct sum of SC and IDSC is 92.16%. As we can see, our proposed SCA leads to the best performance in each part by setting k_1 to 7 and k_2 to 4, and improves the baseline method by a large margin.

D. Natural Image Retrieval

We first demonstrate the performance of SCA for natural image retrieval in Ukbench image dataset [10], which consists of 10,200 images. The whole dataset is divided into 2,550 categories with only 4 images per category. Each image is used in turn as a query. The performance is measured by the average recall of the top four ranked images, referred as N-S score (maximum is 4). The limited ground-truth images per category makes it challenging to achieve good performance for context-based re-ranking methods.

We implement two baseline methods using a local feature and a holistic feature. For local feature, SIFT [34] descriptors are extracted at Hessian-affine [62] interest points, and RootSIFT [22] variant is used. We train the codebook of size 1M using K-means on the independent Flickr60k data [42]. The Bag-of-Words (BoW) feature is computed with hard-assignment and TF-IDF weighting [43]. For holistic feature, we use the 1000-dimensional HSV color histogram ($20\times10\times5$ bins for H, S, V components) following [27], [33]. The HSV histogram is first normalized by its L_1 norm, and finally each element of the histogram is square rooted. The baselines of SIFT and HSV histogram are 3.56 and 3.40 respectively.

We compare our proposed SCA with several state-of-the-art algorithms in Table V. For computing the sparse contextual activation, the size of neighborhood set k_1 is to 4. The parameter for local enhancement enhancement k_2 is set to 2 empirically.

In order to keep the comparison fair, we first compare those methods which use only single feature, but different extra improvements. These selected methods are L_p norm IDF [43], Vector of Aggregated Local Descriptors (VLAD) [49], Triangulation embedding [63], Spatial Contextual Weighting (SCW) [64], Tensor Product Graph (TPG) [17], Burstiness [41], RNN re-ranking [65] and Contextual Dissimilarity Measure [1].

As shown in Table V, our proposed SCA improves the baseline of SIFT from 3.56 to 3.69 and improves the baseline of HSV histogram from 3.40 to 3.56, which makes a significant gain in performance. The performances of Graph Transduction (GT) [3] are also implemented by us using the same features as those of SCA, which are 3.60 for SIFT feature and 3.53 for HSV histogram respectively. Among those methods, GT, TPG, RNN re-ranking and CDM follow the similar principle with our method. They all attach importance to the local context of a given image in the manifold and adopt an iterative strategy to update the similarity measure, but achieve inferior performances compared with SCA, which does not need to iterate until convergence. It can be drawn that SCA can get better performances by exploiting a more reliable distance measure with higher time efficiency.

The performance comparison of the algorithms using multi-feature is also conducted in Table V. We list nearly all the feature fusion algorithms which report their N-S scores to my best knowledge now, including Co-indexing [36], Coupled Binary Embedding [35], Bayes [66], Co-transduction [29], CrDP [67], c-MI [33] and Graph Fusion [27].

Graph Fusion is a representative re-ranking algorithm which deals with multiple input features. The performances of two variants of Graph Fusion, Graph PageRank and Graph Desnity, are 3.76 and 3.77 respectively. By contrast, our proposed feature fusion strategy with SIFT and HSV histogram achieves **3.86 N-S score**. Considering that c-MI [33]^b and Coupled Binary Embedding [35]^b actually fuse three features (fuse SIFT and color descriptor at the indexing level, and HSV histogram descriptor is also combined at the post-processing level), it seems unfair to compare them with SCA, since we actually utilize two features only. However, as can be seen from the table, the performance of SCA is also higher than both of them. Compared with our baseline methods used here, we improve the performance from 3.40 (HSV histogram) and 3.56 (SIFT) to 3.86 significantly. The superior performance demonstrates the discriminative power of our SCA-based feature fusion algorithm.

E. Discussion

Two important parameters are involved in our method: the size of neighborhood set k_1 in sparse contextual activation and the size of neighborhood set k_2 for local consistency enhancement. In this section, we will discuss their effect on the retrieval performance, and compare with the Jaccard re-ranking to show the performance improvement brought by SCA simultaneously.

1) The Parameter k_1: The performance of standard Jaccard distance (Eq. (1)) and the proposed sparse contextual activation (Eq. (7)) with regard to the value of k_1 are reported in Fig. 4. We utilize N-S score for Ukbench dataset and First Tier for PSB dataset to evaluate the retrieval performance.

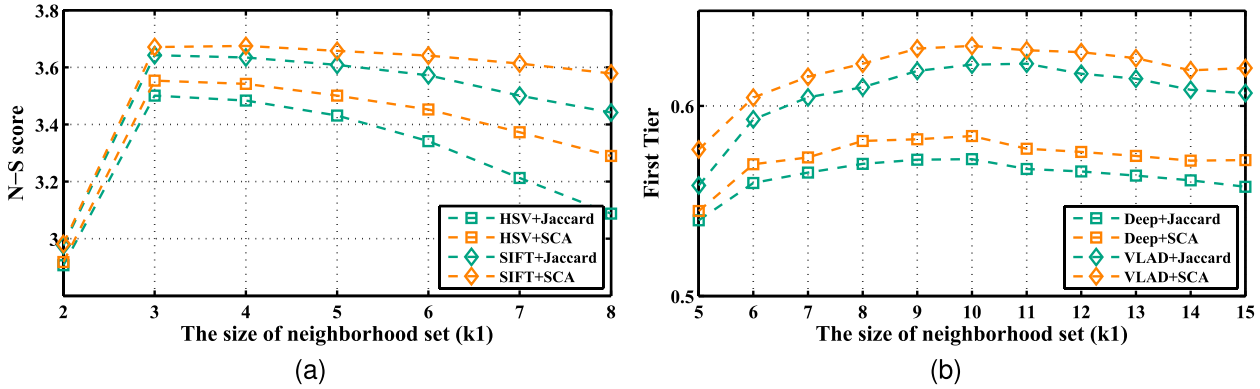


Fig. 4. The impact of the neighborhood set size on retrieval accuracy. N-S score for Ukbench (a), and First Tier for PSB (b) are presented.

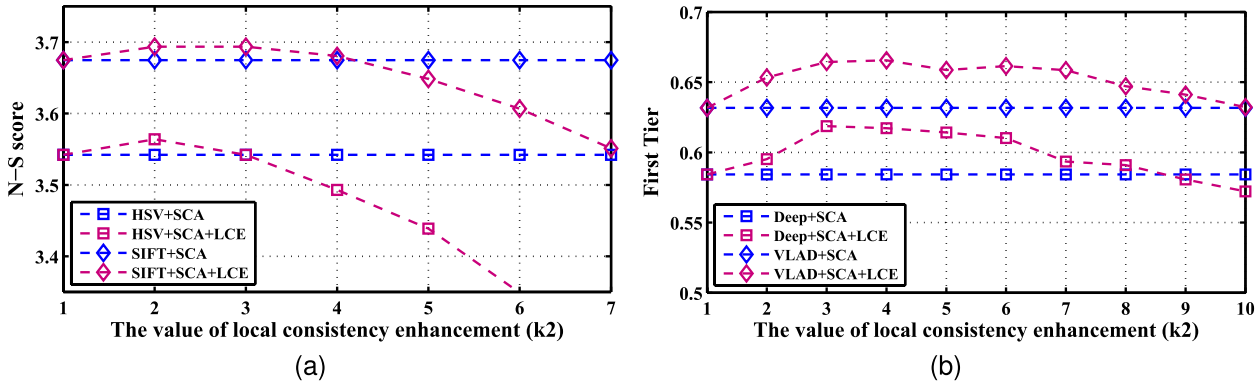


Fig. 5. The impact of the parameter k_2 in local consistency enhancement on retrieval accuracy. N-S score for Ukbench (a) and First Tier for PSB (b) are presented.

It can be observed that SCA outperforms the standard Jaccard distance consistently when the size of neighborhood set k_1 varies, which demonstrates firmly that it is beneficial to bring the weights into account such that the importance of top-ranked images is increased. On the other hand, we can find that when k_1 increase, the performance first increases and drops later after k_1 reaches a certain threshold. It is easy to understand the decreasing of performance when k_1 is relatively larger, owing to the fact that the percentage of false positives will increase in that case.

2) The Parameter k_2: In Fig. 5, we discuss the effect the parameter used in local consistency enhancement by fixing the size of neighborhood set to 4 for Ukbench dataset and 10 for PSB dataset following Section V-D and Section V-A respectively. Note that when k_2 is equal to 1, the local consistency enhancement is not applied at all. Hence, the performances of SCA and SCA with local consistency enhancement are identical at the starting point of the curves in Fig. 5.

It is observed that when local consistency enhancement is applied by using a proper factor k_2, the performance of sparse contextual activation can be improved further. Such a phenomenon demonstrates our previous analysis in Section III-D. It should also be mentioned that if a much larger k_2 is used, the performance will decease without question. Since local consistency enhancement considers the contributions from the neighbors of the second order, it is easier to include the negative effects of noise and outliers compared with only using the direct neighbors (*i.e.*, the neighbors of the first order). However, we find local consistency enhancement of great help to improve the retrieval performance when k_2 is small.

3) Metric: Note that our sparse contextual activation is computed using Eq. (7), a generalized version of Jaccard distance. Indeed, many other metrics can be used here despite the fact that these metrics cannot be interpreted well using the set theory, such as L_r distance,[1] χ^2 distance[2] and Hellinger distance.[3]

In Table VI, we list the performances of sparse contextual activation with LCE or without LCE under different metrics in PSB dataset. The baseline method is the deep feature and First Tier is chosen as the evaluation measure. We can find that the metric defined in Eq. (7) outperforms all the other evaluation metrics. The performance of the widely-used Euclidean distance is not satisfactory enough.

4) Discussion on Rank Aggregation: In Fig. 6, we compare the performances of using low set alone, using high set alone and using the linear combination of high set and low set. First, as the figure shows, the performances of high set and low set reach the peak at different values of k_1, and this value for low

[1]The L_r distance for two SCAs F_q and F_p is $(\sum_i^N |F_{q,i} - F_{p_i}|^r)^{\frac{1}{r}}$.

[2]The χ^2 distance for two SCAs F_q and F_p is $\frac{1}{2}\sum_i^N \frac{(F_{q,i}-F_{p,i})^2}{F_{q,i}+F_{p,i}}$.

[3]The Hellinger distance for two SCAs F_q and F_p is $\sum_i^N (\sqrt{F_{q,i}} - \sqrt{F_{p_i}})^2$.

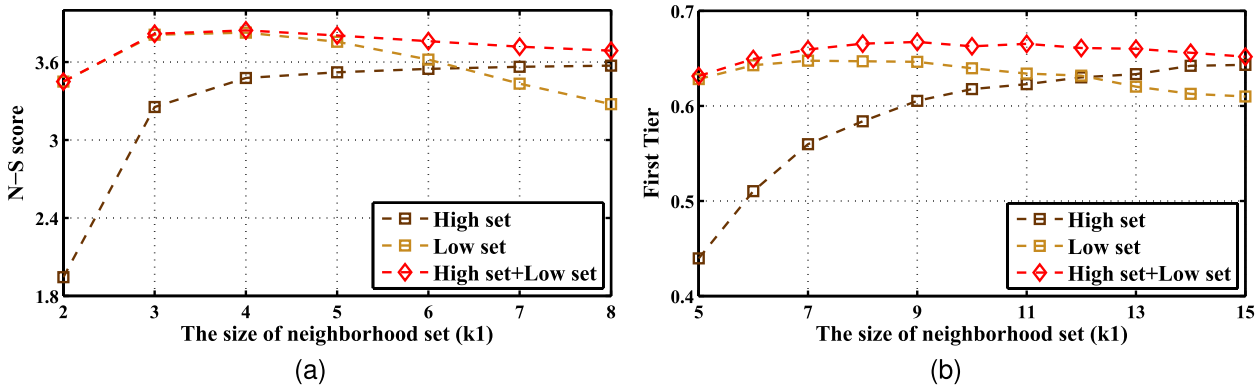


Fig. 6. The discussion about rank aggregation. The performances of high set, low set and their combination are presented. N-S score for Ukbench (a), First Tier for PSB (b) are presented.

TABLE VI
THE PERFORMANCE COMPARISON OF SCA UNDER DIFFERENT METRICS

Metric	Without LCE	With LCE
L_0	0.572	0.550
L_1	0.574	0.610
L_2	0.565	0.583
L_∞	0.387	0.545
χ^2	0.573	0.612
Hellinger	0.573	0.612
ours	**0.584**	**0.617**

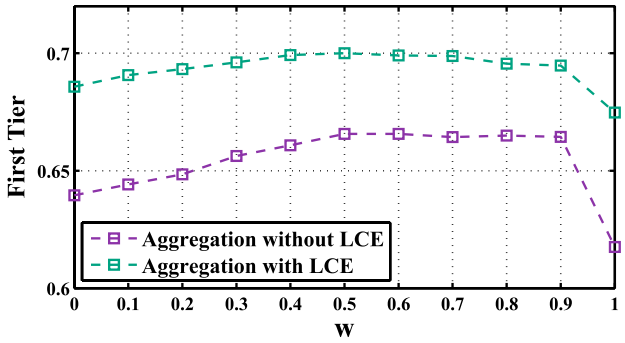


Fig. 7. The influence of weights on PSB dataset.

set is usually smaller than that for high set. Second, the two curves that represents the performance of high set and low set have a intersection point, before which the performance of low set is superior to high set. At last, the linear combination of high set and low set is always better than using either one alone.

5) Feature Weight: In Eq. (18), high set and low set are combined linearly with equal weights for rank aggregation task. To clarify the influence of weight, an experiment on PSB dataset is conducted by setting the weight of high set to w, and that of low set to $1-w$. w ranges from 0 to 1 with a step size of 0.1, and the influence of w on retrieval performance (First Tier) is given in Fig. 7. As the figure shows, a linear combination of high set and low set outperforms their individual counterpart ($w = 0$ or 1), which is consistent to previous analysis. The best performance is achieved when w is around 0.5. However, the performance only changes a little or remains the same when w shifts to 0.4 or 0.6. Hence, we set the default value of w to 0.5 for simplicity. Note that one can learn the weights automatically if some prior information (*e.g.* the baseline performances) is known.

F. Complexity Analysis

Considering that SCA is a post-processing procedure that focuses on re-ranking instead of ranking initialization, we only analyse the time efficiency and memory cost after the original ranking are given. The operation in the pre-processing, such as feature extraction, feature quantization, pairwise distance computation, *etc*., are beyond the scope of our concern.

1) Time Complexity: Given an image collection with N elements and their pairwise distances known, the computation of the sparse contextual activation for an given image x_q requires the computational complexity $O(k_1)$, where k_1 is the size of the neighborhood set. When local consistency enhancement determined by the parameter k_2 is applied, the computation time cost for SCA is equivalent to $O(k_2 \times k_1)$. We can find that the initiation of SCA is independent of the image collection size N, but related to the size of the local distribution around x_q on the manifold. Usually, the values of k_1 and k_2 are much smaller than N, which indicates that the computation of SCA for a single image can be assumed to achieve in $O(1)$ efficiently.

As for re-ranking with SCA, the direct computation of the distance between x_q and all the other images in the database requires $O(2 \times N^2)$. However, the time complexity can be reduced significantly by using inverted index as presented in Section III-C. Embedded with inverted index, the MIN operation needs $O(M \times k_1)$ through Eq. (11) in the worst case, where M denotes the number of images that have overlaps in neighborhood sets with x_q. MAX only needs $O(M)$ through Eq. (12). In summary, the time complexity is reduced from $O(2 \times N^2)$ to $O(M \times (k_1 + 1))$ by introducing the inverted index. In fact, the value of M is also much smaller than N (*e.g.*, for the HSV histogram in Ukbench dataset, the average value of M is merely 7.46).

All our experiments are carried out on a desktop machine with an Intel(R) Core(TM) i5 CPU (3.40 GHz) and

TABLE VII
THE COMPARISON OF AVERAGE RE-RANKING TIME

Methods	Jaccard	TPG [17]	LCDP [8]	SCA
Ukbench	3.22s	3.81s	35.12ms	**0.39*ms***
PSB	0.28s	75.27ms	12.22ms	**0.17*ms***
MEPG-7	0.44s	0.26s	24.89ms	**0.20*ms***

16 GB memory. In Ukbench dataset, the generation of SCA takes 0.14ms and local consistency enhancement takes 0.36ms per image. It takes 0.33s to build the inverted index for the whole dataset. For a given query, the average cost for re-ranking is only **0.39*ms***.

Many previous re-ranking algorithms (*e.g.* [8], [17], [29]) are time-consuming due to their high time complexity ($O(N^3)$ in most cases). They usually need an iterative procedure to spread the affinity values on a huge graph which establishes the relationship among all the images. Table VII compares the average time for re-ranking of some typical algorithms, including Jaccard re-ranking defined in Eq. (1). As the table shows, SCA reduces the time cost of Jaccard re-ranking by more than 8,000 times in Ukbench dataset when inverted index is applied. Compared with the two representative diffusion-based re-ranking methods, SCA is 90 times faster than LCDP, and 9,800 times faster than TPG. It indicates that our proposed re-ranking method has the potential for large scale re-ranking task.

2) Space Complexity: The scale of the inverted index is equivalent to the number of images in the database, which indicates the space complexity of SCA is $O(N)$.

For each entry in the inverted index used for re-ranking, we use 4 bytes to store one image index. 4 bytes (single format) are used to denote the activation for a certain image. On the Ukbench dataset, it only takes 65.2KB to save the inverted index in our implementation. Considering the big improvement on the computational efficiency, the minor extra memory cost can be tolerated.

VI. CONCLUSION AND FUTURE WORK

In this paper, we propose an extremely efficient algorithm called Sparse Contextual Activation (SCA) for visual re-ranking. SCA is a merely single vector that encodes the contextual distribution for an image. The re-ranking procedure can be simply conducted by vector comparison using generalized Jaccard distance. For the first time, inverted index is introduced into re-ranking task, which makes is possible that re-ranking for a single query can be finished within one millisecond. Local Consistency Enhancement (LCE) is also developed to improve the performance of SCA further. The experimental results on PSB dataset (3D object), WM-SHREC07 (3D competition), YALE dataset B (face), MPEG-7 dataset (shape) and Ukbench dataset (natural image) demonstrate the effectiveness and efficiency of the proposed method.

Note that we design an inverted index for visual re-ranking, so an efficient method about the dynamical of insertion, deletion and modification should be taken into consideration. Meanwhile, since SCA is defined in the context, it requires for accurate contextual distribution if possible. So does it benefit from adding more extra images or artificial ghost points [68] to densify the feature space? Moreover, as SCA provides a more efficient and effective similarity measure similar to diffusion process, it can be potentially applied to other important tasks such as image segmentation [11], [12], [69], image matching/ registration [70], [71], visual object tracking [13], and analysis of medical data [72]. We leave all these problems for the future work.

REFERENCES

[1] H. Jegou, C. Schmid, H. Harzallah, and J. Verbeek, "Accurate image search using the contextual dissimilarity measure," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 1, pp. 2–11, Jan. 2010.

[2] M. Donoser and H. Bischof, "Diffusion processes for retrieval revisited," in *Proc. CVPR*, Jun. 2013, pp. 1320–1327.

[3] X. Bai, X. Yang, L. J. Latecki, W. Liu, and Z. Tu, "Learning context-sensitive shape similarity by graph transduction," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 5, pp. 861–874, May 2010.

[4] J. Sivic and A. Zisserman, "Video Google: A text retrieval approach to object matching in videos," in *Proc. ICCV*, Oct. 2003, pp. 1470–1477.

[5] J. Wang, Y. Li, X. Bai, Y. Zhang, C. Wang, and N. Tang, "Learning context-sensitive similarity by shortest path propagation," *Pattern Recognit.*, vol. 44, nos. 10–11, pp. 2367–2374, Oct./Nov. 2011.

[6] D. C. G. Pedronette, J. Almeida, and R. da S. Torres, "A scalable re-ranking method for content-based image retrieval," *Inf. Sci.*, vol. 265, pp. 91–104, May 2014.

[7] L. Luo, C. Shen, C. Zhang, and A. van den Hengel, "Shape similarity analysis by self-tuning locally constrained mixed-diffusion," *IEEE Trans. Multimedia*, vol. 15, no. 5, pp. 1174–1183, Aug. 2013.

[8] X. Yang, S. Koknar-Tezel, and L. J. Latecki, "Locally constrained diffusion process on locally densified distance spaces with applications to shape retrieval," in *Proc. CVPR*, Jun. 2009, pp. 357–364.

[9] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Object retrieval with large vocabularies and fast spatial matching," in *Proc. CVPR*, Jun. 2007, pp. 1–8.

[10] D. Nistér and H. Stewénius, "Scalable recognition with a vocabulary tree," in *Proc. CVPR*, 2006, pp. 2161–2168.

[11] B. Wang and Z. Tu, "Affinity learning via self-diffusion for image segmentation and clustering," in *Proc. CVPR*, Jun. 2012, pp. 2312–2319.

[12] X. Wang, Y. Tang, S. Masnou, and L. Chen, "A global/local affinity graph for image segmentation," *IEEE Trans. Image Process.*, vol. 24, no. 4, pp. 1399–1411, Apr. 2015.

[13] Y. Zhou, X. Bai, W. Liu, and L. J. Latecki, "Fusion with diffusion for robust visual tracking," in *Proc. NIPS*, 2012, pp. 2987–2995.

[14] P. Kontschieder, M. Donoser, and H. Bischof, "Beyond pairwise shape similarity analysis," in *Proc. ACCV*, 2009, pp. 655–666.

[15] A. Egozi, Y. Keller, and H. Guterman, "Improving shape retrieval by spectral matching and meta similarity," *IEEE Trans. Image Process.*, vol. 19, no. 5, pp. 1319–1327, May 2010.

[16] J. J.-Y. Wang and Y. Sun, "From one graph to many: Ensemble transduction for content-based database retrieval," *Knowl.-Based Syst.*, vol. 65, pp. 31–37, Jul. 2014.

[17] X. Yang, L. Prasad, and L. J. Latecki, "Affinity learning with diffusion on tensor product graph," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 28–38, Jan. 2013.

[18] Y. Chen, X. Li, A. Dick, and R. Hill, "Ranking consistency for image matching and object retrieval," *Pattern Recognit.*, vol. 47, no. 3, pp. 1349–1360, Mar. 2014.

[19] D. C. G. Pedronette and R. da S. Torres, "Image re-ranking and rank aggregation based on similarity of ranked lists," *Pattern Recognit.*, vol. 46, no. 8, pp. 2350–2360, Aug. 2013.

[20] D. C. G. Pedronette, O. A. B. Penatti, and R. da S. Torres, "Unsupervised manifold learning using reciprocal kNN Graphs in image re-ranking and rank aggregation tasks," *Image Vis. Comput.*, vol. 32, no. 2, pp. 120–130, Feb. 2014.

[21] O. Chum, A. Mikulik, M. Perdoch, and J. Matas, "Total recall II: Query expansion revisited," in *Proc. CVPR*, Jun. 2011, pp. 889–896.

[22] R. Arandjelovic and A. Zisserman, "Three things everyone should know to improve object retrieval," in *Proc. CVPR*, Jun. 2012, pp. 2911–2918.

[23] R. Fagin, R. Kumar, and D. Sivakumar, "Comparing top k lists," *SIAM J. Discrete Math.*, vol. 17, no. 1, pp. 134–160, 2003.

[24] W. Webber, A. Moffat, and J. Zobel, "A similarity measure for indefinite rankings," *ACM Trans. Inf. Syst.*, vol. 28, no. 4, p. 20, Nov. 2010.

[25] R. R. Coifman and S. Lafon, "Diffusion maps," *Appl. Comput. Harmon. Anal.*, vol. 21, no. 1, pp. 5–30, Jul. 2006.

[26] P. V. Gehler and S. Nowozin, "On feature combination for multiclass object classification," in *Proc. ICCV*, Sep./Oct. 2009, pp. 221–228.

[27] S. Zhang, M. Yang, T. Cour, K. Yu, and D. N. Metaxas, "Query specific fusion for image retrieval," in *Proc. ECCV*, 2012, pp. 660–673.

[28] S. Zhang, M. Yang, T. Cour, K. Yu, and D. N. Metaxas, "Query specific rank fusion for image retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 4, pp. 803–815, Apr. 2015.

[29] X. Bai, B. Wang, C. Yao, W. Liu, and Z. Tu, "Co-transduction for shape retrieval," *IEEE Trans. Image Process.*, vol. 21, no. 5, pp. 2747–2757, May 2012.

[30] Z.-H. Zhou and M. Li, "Tri-training: Exploiting unlabeled data using three classifiers," *IEEE Trans. Knowl. Data Eng.*, vol. 17, no. 11, pp. 1529–1541, Nov. 2005.

[31] Z.-H. Zhou and M. Li, "Semi-supervised regression with Co-training," in *Proc. IJCAI*, 2005, pp. 908–913.

[32] S. Zhang, J. Huang, H. Li, and D. N. Metaxas, "Automatic image annotation and retrieval using group sparsity," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 42, no. 3, pp. 838–849, Jun. 2012.

[33] L. Zheng, S. Wang, Z. Liu, and Q. Tian, "Packing and padding: Coupled multi-index for accurate image retrieval," in *Proc. CVPR*, Jun. 2014, pp. 1947–1954.

[34] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

[35] L. Zheng, S. Wang, and Q. Tian, "Coupled binary embedding for large-scale image retrieval," *IEEE Trans. Image Process.*, vol. 23, no. 8, pp. 3368–3380, Aug. 2014.

[36] S. Zhang, M. Yang, X. Wang, Y. Lin, and Q. Tian, "Semantic-aware Co-indexing for image retrieval," in *Proc. ICCV*, Dec. 2013, pp. 1673–1680.

[37] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, pp. 2323–2326, Dec. 2000.

[38] S. Belongie, J. Malik, and J. Puzicha, "Shape matching and object recognition using shape contexts," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 4, pp. 509–522, Apr. 2002.

[39] H. Ling and D. W. Jacobs, "Shape classification using the inner-distance," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 2, pp. 286–299, Feb. 2007.

[40] J. Wang, X. Bai, X. You, W. Liu, and L. J. Latecki, "Shape matching and classification using height functions," *Pattern Recognit. Lett.*, vol. 33, no. 2, pp. 134–143, 2012.

[41] H. Jegou, M. Douze, and C. Schmid, "On the burstiness of visual elements," in *Proc. CVPR*, Jun. 2009, pp. 1169–1176.

[42] H. Jégou, M. Douze, and C. Schmid, "Hamming embedding and weak geometric consistency for large scale image search," in *Proc. ECCV*, 2008, pp. 304–317.

[43] L. Zheng, S. Wang, and Q. Tian, "L_p-norm IDF for scalable image retrieval," *IEEE Trans. Image Process.*, vol. 23, no. 8, pp. 3604–3617, Aug. 2014.

[44] P. Papadakis, I. Pratikakis, T. Theoharis, and S. J. Perantonis, "PANORAMA: A 3D shape descriptor based on panoramic views for unsupervised 3D object retrieval," *Int. J. Comput. Vis.*, vol. 89, nos. 2–3, pp. 177–192, Sep. 2010.

[45] P. Shilane, P. Min, M. M. Kazhdan, and T. A. Funkhouser, "The princeton shape benchmark," in *Proc. SMI*, Jun. 2004, pp. 167–178.

[46] D. Giorgi, S. Biasotti, and L. Paraboschi, "Shape retrieval contest 2007: Watertight models track," in *Proc. SHREC Competition*, 2007, pp. 1–8.

[47] A. S. Georghiades, P. N. Belhumeur, and D. Kriegman, "From few to many: Illumination cone models for face recognition under variable lighting and pose," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 6, pp. 643–660, Jun. 2001.

[48] L. J. Latecki, R. Lakamper, and T. Eckhardt, "Shape descriptors for non-rigid shapes with a single closed contour," in *Proc. CVPR*, Jun. 2000, pp. 424–429.

[49] H. Jégou, F. Perronnin, M. Douze, J. Sánchez, P. Pérez, and C. Schmid, "Aggregating local image descriptors into compact codes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 9, pp. 1704–1716, Sep. 2012.

[50] H. Tabia, D. Picard, H. Laga, and P.-H. Gosselin, "Compact vectors of locally aggregated tensors for 3D shape retrieval," in *Proc. 3DOR*, 2013, pp. 17–24.

[51] S. Bai, X. Bai, W. Liu, and F. Roli, "Neural shape codes for 3D model retrieval," *Pattern Recognit. Lett.*, vol. 65, pp. 15–21, Nov. 2015.

[52] D.-Y. Chen, X.-P. Tian, Y.-T. Shen, and M. Ouhyoung, "On visual similarity based 3D model retrieval," *Comput. Graph. Forum*, vol. 22, no. 3, pp. 223–232, Sep. 2003.

[53] H. Tabia, M. Daoudi, J.-P. Vandeborre, and O. Colot, "A new 3D-matching method of nonrigid and partially similar models using curve analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 4, pp. 852–858, Apr. 2011.

[54] D. V. Vranic, "DESIRE: A composite 3D-shape descriptor," in *Proc. ICME*, Jul. 2005, pp. 962–965.

[55] M. Liu, B. C. Vemuri, S.-I. Amari, and F. Nielsen, "Shape retrieval using hierarchical total Bregman soft clustering," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 12, pp. 2407–2419, Dec. 2012.

[56] H. Tabia, H. Laga, D. Picard, and P.-H. Gosselin, "Covariance descriptors for 3D shape matching and retrieval," in *Proc. CVPR*, Jun. 2014, pp. 4185–4192.

[57] P. Papadakis, I. Pratikakis, T. Theoharis, G. Passalis, and S. J. Perantonis, "3D object retrieval using an efficient and compact hybrid shape descriptor," in *Proc. 3DOR*, 2008, pp. 9–16.

[58] X. Bai, C. Rao, and X. Wang, "Shape vocabulary: A robust and efficient shape representation for shape matching," *IEEE Trans. Image Process.*, vol. 23, no. 9, pp. 3935–3949, Sep. 2014.

[59] T. Furuya and R. Ohbuchi, "Fusing multiple features for shape-based 3d model retrieval," in *Proc. BMVC*, 2014, pp. 1–12.

[60] T. Ojala, M. Pietikäinen, and T. Mäenpää, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 971–987, Jul. 2002.

[61] D. C. G. Pedronette and R. da S. Torres, "Exploiting pairwise recommendation and clustering strategies for image re-ranking," *Inf. Sci.*, vol. 207, pp. 19–34, Nov. 2012.

[62] K. Mikolajczyk and C. Schmid, "Scale & affine invariant interest point detectors," *Int. J. Comput. Vis.*, vol. 60, no. 1, pp. 63–86, Oct. 2004.

[63] H. Jégou and A. Zisserman, "Triangulation embedding and democratic aggregation for image search," in *Proc. CVPR*, Jun. 2014, pp. 3310–3317.

[64] X. Wang, M. Yang, T. Cour, S. Zhu, K. Yu, and T. X. Han, "Contextual weighting for vocabulary tree based image retrieval," in *Proc. ICCV*, Nov. 2011, pp. 209–216.

[65] D. Qin, S. Gammeter, L. Bossard, T. Quack, and L. Van Gool, "Hello neighbor: Accurate object retrieval with k-reciprocal nearest neighbors," in *Proc. CVPR*, Jun. 2011, pp. 777–784.

[66] L. Zheng, S. Wang, W. Zhou, and Q. Tian, "Bayes merging of multiple vocabularies for scalable image retrieval," in *Proc. CVPR*, Jun. 2014, pp. 1963–1970.

[67] B. Wang, J. Jiang, W. Wang, Z.-H. Zhou, and Z. Tu, "Unsupervised metric fusion by cross diffusion," in *Proc. CVPR*, Jun. 2012, pp. 2997–3004.

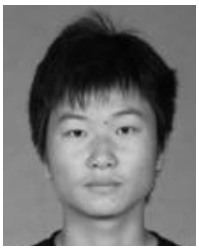
[68] X. Yang, X. Bai, S. Köknar-Tezel, and L. J. Latecki, "Densifying distance spaces for shape and image retrieval," *J. Math. Imag. Vis.*, vol. 46, no. 1, pp. 12–28, May 2013.

[69] C. Li, C. Xu, C. Gui, and M. D. Fox, "Distance regularized level set evolution and its application to image segmentation," *IEEE Trans. Image Process.*, vol. 19, no. 12, pp. 3243–3254, Dec. 2010.

[70] J. Ma, J. Zhao, J. Zhao, J. Tian, A. Yuille, and Z. Tu, "Robust point matching via vector field consensus," *IEEE Trans. Image Process.*, vol. 23, no. 4, pp. 1706–1721, Apr. 2014.

[71] J. Ma, W. Qiu, J. Zhao, Y. Ma, A. L. Yuille, and Z. Tu, "Robust L_2E estimation of transformation for non-rigid registration," *IEEE Trans. Signal Process.*, vol. 63, no. 5, pp. 1115–1129, Mar. 2015.

[72] B. Wang *et al.*, "Similarity network fusion for aggregating data types on a genomic scale," *Nature Methods*, vol. 11, no. 3, pp. 333–337, Jan. 2014.



Song Bai (S'01) received the B.S. degree in electronics and information engineering from the Huazhong University of Science and Technology, Wuhan, China, in 2013, where he is currently pursuing the Ph.D. degree with the School of Electronic Information and Communications. His research interests include shape analysis, image classification, and retrieval.

Xiang Bai (SM'07) received the B.S., M.S., and Ph.D. degrees from the Huazhong University of Science and Technology (HUST), Wuhan, China, in 2003, 2005, and 2009, respectively, all in electronics and information engineering. He is currently a Professor with the School of Electronic Information and Communications, HUST. He is also the Vice Director of the National Center of Anti-Counterfeiting Technology with HUST. His research interests include object recognition, shape analysis, scene text recognition, and intelligent systems.