Holons Visual Representation for Image Retrieval

Le Dong, Yan Liang, Gaipeng Kong, Qianni Zhang, Xiaochun Cao, and Ebroul Izquierdo

***Abstract*—Along with the enlargement of image scale, convolutional local features, such as SIFT, are ineffective for representing or indexing and more compact visual representations are required. Due to the intrinsic mechanism, the state-of-the-art vector of locally aggregated descriptors (VLAD) has a few limits. Based on this, we propose a new descriptor named holons visual representation (HVR). The proposed HVR is a derivative mutational self-contained combination of global and local information. It exploits both global characteristics and the statistic information of local descriptors in the image dataset. It also takes advantages of local features of each image and computes their distribution with respect to the entire local descriptor space. Accordingly, the HVR is computed by a two-layer hierarchical scheme, which splits the local feature space and obtains raw partitions, as well as the corresponding refined partitions. Then, according to the distances from the centroids of partition spaces to local features and their spatial correlation, we assign the local features into their nearest raw partitions and refined partitions to obtain the global description of an image. Compared with VLAD, HVR holds critical structure information and enhances the discriminative power of individual representation with a small amount of computation cost, while using the same memory overhead. Extensive experiments on several benchmark datasets demonstrate that the proposed HVR outperforms conventional approaches in terms of scalability as well as retrieval accuracy for images with similar intra local information.**

***Index Terms*—Clustering, image retrieval, similarity distance regularization, visual representation.**

I. Introduction

WE are interested in the problem of accurately and efficiently finding the most similar images of a given object or scene from the image databases. Along with the emergence of large-scale image repositories, image retrieval faces more challenges. First, the extraction of image representation and search for similar images lead to huge computation overheads and memory consumption, which are stretching beyond the capacity of even the advanced computer systems. Second, different with the medium-scale image retrieval, the large-scale counterparts impose a crucial additional challenge due to the fact that retrieval performance usually decreases proportionally to the size of image database. As a consequence, the scale and image diversity of modern databases inevitably bring a critical need for compact visual representation with high discriminative ability.

Numerous efforts have been made to counter this challenge, among which the classic bag-of-words (BOW) model is the most widely adopted method and has achieved great success in image retrieval [1]. Inspired by the relatively mature text retrieval techniques, BOW model trains numerous distinctive visual words and describes an image using a high-dimensional histogram with respect to the visual words, which is beneficial for both the effectiveness and efficiency of image retrieval. BOW model usually employs TF-IDF [2] to weigh visual words, where the TF has the capability of indicating the importance of visual words in a query image and the IDF is able to reflect the discriminative abilities of visual words in image database. However, in the BOW-based retrieval algorithms, each image feature is indexed individually as an item in the inverted index structure. Thus, the memory cost per image for indexing is proportionally to the feature number. Since an image usually contains thousands of local features, the memory overhead to index large-scale image dataset is extremely heavy, which limits the retrieval scalability in real applications.

In addition, Jegou *et al.* proposed vector of locally aggregated descriptors (VLAD) [7], which has attached much attention and achieved a more scalable image retrieval. VLAD records the sum of the differences between local features and cluster centroids of image datasets. Through dimensionality reduction and compression, the obtained VLAD can be reduced to a very compact vector. Therefore, VLAD can obtain better retrieval performance than BOW with less memory consumption of image representations in the index files. Unfortunately, VLAD is limited in real applications to some extent due to its intrinsic mechanism. When two local descriptors are assigned to the same visual word, they may have the same absolute value but the adverse signs. In this case, their residual vectors cancel out each other. As illustrated in Fig. 1, the sum of three residual vectors in the left visual word region is equal to the sum of only one residual vector in the right region. Thus, VLAD can not distinguish the sums of the two residual vectors, which will significantly affect the descriptive ability of VLAD. Moreover, VLAD only uses the local descriptors and omits other useful information contained in the original local features such as the spatial information.

Manuscript received September 24, 2015; revised January 14, 2016; accepted January 29, 2016. Date of publication March 1, 2016; date of current version March 15, 2016. This work was supported in part by the National Natural Science Foundation of China under Grant 61370149, in part by the Fundamental Research Funds for the Central Universities under Grant ZYGX2013J083, by the Scientific Research Foundation for the Returned Overseas Chinese Scholars, State Education Ministry under Grant LXHG42DL, by the National Basic Research Program of China under Grant 2013CB329305, by the National Natural Science Foundation of China under Grant 61422213, and by the Strategic Priority Research Program of the Chinese Academy of Sciences. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Changsheng Xu.

L. Dong and G. Kong are with the School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China (e-mail: ledong@uestc.edu.cn; konggaipeng@163.com).

Y. Liang is with the the Institute of Computer Application, China Academy of Engineering Physics, Mianyang 621900, China (e-mail: xuehuaiyan@163.com).

Q. Zhang and E. Izquierdo are with the School of Electronic Engineering and Computer Science, Queen Mary University of London, London E1 4NS, U.K. (e-mail: qianni.zhang@qmul.ac.uk; ebroul.izquierdo@qmul.ac.uk).

X. Cao is with the State Key Laboratory of Information Security, Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100093, China, and also with the School of Computer Science and Technology, Tianjin University, Tianjin 300072, China (e-mail: caoxiaochun@iie.ac.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/TMM.2016.2530399

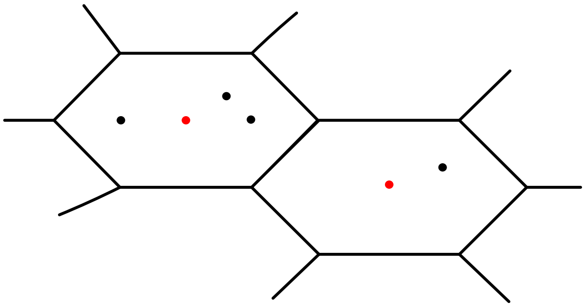


Fig. 1. Different samples with the same residual vector.

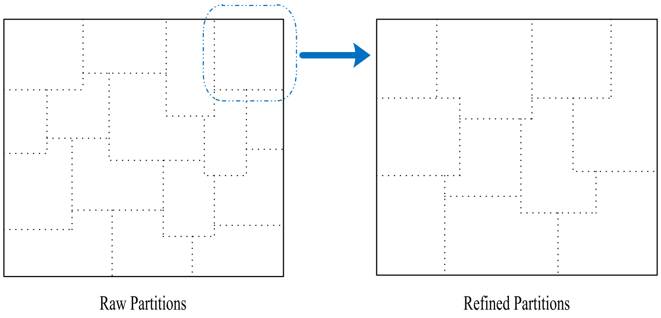


Fig. 2. Two-layer hierarchical local feature space split scheme. Here, the local feature space is first split into 16 raw partitions, and then each raw partition is split into 10 refined partitions.

In this paper, we aim at designing a compact visual representation of an image and improving the discriminative ability of image representation by considering more clues. To achieve this goal, we propose a new image description, called holons visual representation (HVR). It includes two visual representations, distance histogram based clustering (DHC) and distance and orientation histogram based clustering (DOHC), whose idea differs from BOW and VLAD. HVR is a derivative mutational self-contained combination of global and local information. Different from the traditional methods [16], [36], which learn a linear transformation to integrate the global and local features of an image, we utilize both the global characteristics of the image datasets and the local visual information of an image. In HVR, the local features, such as SIFT, are utilized to denote the local visual information of an image. The partitions and their centroids of the whole local feature space are trained to describe the global characteristics of the image datasets. As illustrated in Fig. 2, HVR employs a two-layer hierarchical scheme to extract the global information and statistics of the local feature set of image datasets. We first roughly split the whole local feature space into K raw partitions by clustering, and then each raw partition is split into N refined partitions using the same method.

To obtain the HVR of an image, we first compute the distances between SIFT descriptors and the centroids of the raw partitions, and assign the SIFT features into the nearest raw partitions. Then, through the similar method, the SIFT features belonging to each raw partition are assigned into their nearest refined partitions according to the descriptor distance and the spatial consistency of the local features. The value of each dimension in HVR describes the number of SIFT features assigned into the refined partition. Therefore, HVR utilizes the framework of assigning the SIFT features into the corresponding partitions and aggregating them, which effectively combines two useful information, namely, the global information of the whole feature space of image datasets and the SIFT features of the input image. More specifically, HVR describes the global distribution of SIFT features of an image with respect to the whole feature space of image datasets.

Different from the traditional BOW, HVR employs a two-layer hierarchical scheme to extract the global information and statistics of the local feature set of image datasets. At the same time, there are clear distinctions between HVR and vocabulary tree or hierarchical BOW. In fact the two methods are based on totally different ideas. Firstly, HVR is a compact image representation obtained by aggregating SIFT features, while BOW is a high-dimensional descriptor without aggregation. Next, in hierarchical BOW, the visual words are the clustering centers of SIFT descriptors and used during the whole process. In HVR, there are two different codebooks. The first is the centroids of trained partitions of SIFT feature space and utilized to form HVRs. The second is the clustering centers of the whole and sub HVR vectors, and is used in quantization, index, and retrieval. Moreover, the number of refined partitions is much less than that of visual words in hierarchical BOW. This is because HVR just needs a general holistic statistics of SIFT feature space, while BOW needs a detailed division which has a considerable influence on the quantization and similarity measure. Finally, since the refined partitions belonging to the same raw partition are associate with the same raw centroid, these can be used as valuable correspondence information among the N dimensions of the same raw partition in HVR. Hence, the dimensions in the same raw partition are bundled together to serve as a component to build indexes and retrieve. This is different from hierarchical BOW representation, in which each dimension is relatively independent and regarded as a component.

In the experiments, we give a detailed analysis of the performance of our proposed HVR. We employ the approximate nearest neighbor search method [9] to build the index files and compute the similarity scores between the query image and database images. We firstly measure the tradeoff between the different parameters of HVR and the corresponding retrieval accuracy, and then make a comparison between the DOHC and HVR. The experimental results show that the proposed HVR enhances the retrieval accuracy on Oxford5k Building and Paris6k datasets. Moreover, considering that the scalability of image representations becomes increasingly important along with the increase of image dataset size, we also evaluate the scalability of HVR to demonstrate the validity of our method when applied to larger scale datasets. 1.26 million images covering 1000 categories of objects from a large-scale image dataset are employed as distractors, which are merged with other medium-scale image datasets in the large-scale image retrieval experiments. The experimental results show that, compared with VLAD, the proposed HVR is with great scalability.

Generally, the *contribution* of this paper can be concluded in three folds.

1) First, the proposed approach addresses the retrieval of the images with the similar intra local structures, rather than the general image retrieval.

2) Second, we propose a brand new HVR, which is a derivative mutational self-contained combination of global and local information. It can overcome the limitations of VLAD and enhance the retrieval accuracy for the images containing similar intra visual information.
3) Last but not the least, for the proposed HVR, we also propose a new normalization scheme, i.e. the optimized raw partition normalization, which integrates signs square root (SSR) and sub-normalization and can significantly decrease the burstiness problem of HVR.

The remainder of the paper is organized as follows: We briefly discuss the related work in Section II. The proposed HVR is elaborated in Section III. Section IV first introduces the data sets and evaluation protocol, as well as the implementation details in experiments. Then a detailed analysis of the experimental results on both medium-scale and large-scale image datasets is made. Finally, the paper is concluded in Section V.

II. Related Work

For image retrieval, how to describe and represent images is a crucial and challenging task. The large-scale image retrieval systems [2], [4]–[6], [11], [17], [23] have been significantly advanced by the introduction of local SIFT descriptor [3] and BOW model which relies on the quantization of local descriptors into visual words. Recently, numerous methods combining multiple types of useful information were proposed to represent the images. For instance, to form a highly discriminative image representation, geometrical relations among local features [11]–[14], [44] and spatial distributions [17]–[19], [42] were utilized; multifarious visual features were bundled to construct reliable high-dimensional features in [15], [34]. The combination of the low-level features was proposed to represent images in [35], [41], [45]. To improve the retrieval accuracy, an intuitive way of image retrieval was presented in [20], [21], [43], in which users can describe the intended search targets with understandable attributes. In addition, post-processing techniques such as reranking [37] and query expansion [24], [38] can boost the accuracy. In this paper, our method focuses on both retrieval accuracy and memory overhead. However, the spatial verification or query expansion are not taken into account, since they are much slower than the retrieval process.

These aforementioned techniques increase the discriminative ability for the image representations and obtain the improvements on retrieval accuracy. However, most of these methods require a significant amount of memory overhead per image, and induce considerable computation overhead to form each image representation as the size of image datasets reaches to a million scale. To release the efficiency and memory constraints, minhash approach [11], [12] was proposed to gain binary vectors of image representation. Nevertheless, compared with BOW representation, the technique reduces the retrieval accuracy and still requires tremendous memory per image. In [39], compact feature based clustering was proposed to represent images by aggregating clustering centers and statistics of SIFT features assigned to each clusters. Although it reduces the computation and memory footprint to some extent, it still needs several feature vectors to describe images. In addition, feature selection was one main technique for dimensionality reduction that involved identifying a subset of the most useful features. Li *et al.* [47], [48] proposed a unsupervised feature selection scheme by by integrating cluster analysis and sparse structural analysis into a joint framework.

The state-of-the-art description encoding methods introduced a new way to form image description, and achieved a higher retrieval accuracy [7], [8], [25]–[27], [29]. Li *et al.* [46] learned a robust representation for images. The proposed framework was a general one which can leverage several well-known algorithms as special cases and elucidate their intrinsic relationships. VLAD [7] and Fisher vector [26], [27] represented images via a highly discriminative vector which was a high-dimensional code. VLAD was based on Fisher kernels, and in fact, it was a simplified nonprobabilistic version of the Fisher kernels. It employed the dimension reduction and compression algorithms to jointly optimize, and used the Kmeans clustering instead of the GMM clustering. VLAD with the same size significantly outperforms the BOW representation. Arandjelovic and Zisserman proposed an intra-normalization method to address the burstiness problem of VLAD, and extracted multiple VLADs for an image to improve the retrieval accuracy [8]. In [25], residual normalization was proposed to make each local descriptor contribute equally to the VLAD. The methods were based on VLAD and used the sum of the residual vectors to describe images. Since the residual vectors may be positive or negative values and can offset, they can not distinguish the different images when the residual vectors of their local descriptors were offset to be the same. Furthermore, they only use the local descriptors and omit other useful information. Based on the above analysis, we propose a brand new image representation, i.e., HVR. We divide the entire local descriptor space of the image datasets into two-layer hierarchical partitions, according to the distances and spacial consistency among the local features of an image and the centroids of the refined partitions. HVR records the distribution of the local features with respect to the entire local features set of the image datasets.

Quantification and the retrieval algorithms are key factors affecting the retrieval efficiency. The vocabulary tree [2] typically contains millions of leaf nodes that represent visual words. Using an inverted file index of visual words, it can avoid directly storing and comparing high-dimensional local descriptors, and reduce the number of target images since only those images sharing the same visual words with the query image needed to be matched. The work in [22] optimized the quantification by softly assigning descriptors to multiple visual words. Some researchers improved the retrieval accuracy of vocabulary tree by taking advantages of more valuable visual information. For instance, a context dissimilarity measure of visual word vectors [30], [31] was learned and a projection from descriptor space [32] was trained to form a new Euclidean space for descriptor comparison and quantization; descriptor contextual weighting and spatial contextual weighting were added in [6]. These methods needed a large amount of visual words, which made the training difficult when the number reaches to millions. In [9], an inverted file system with the asymmetric distance computation (IVFADC) was proposed to avoid learning millions of visual words. In this paper, we employ the inverted file system with asymmetric distance computation to build the index files and compute the similarity.

III. Discriminative HVR

In this section, we first briefly describe the VLAD, and then introduce the proposed HVR, including two different approaches, i.e., (I) DHC and (II) DOHC. The proposed method is based on the scheme of aggregating local features in each partition of the local feature space.

A. Description of VLAD

VLAD is an encoding method for local descriptors, which can transform a variable-size local descriptor set extracted from an image $I = \lceil x_1, x_2, \ldots, x_M \rceil \in R^{D \times M}$ into a fixed size image representation. The input set contains M local descriptors with D dimensions, and the value of D is 128 when SIFT descriptors are extracted to serve as the primitive image representation.

Similar to BOW, a codebook $C = \lceil \mu_1, \ldots, \mu_K \rceil$ is learned on the independent SIFT descriptor samples in advance by clustering, and each cluster center acts as a visual word. The difference from the BOW representation is that the number of visual words K is typically a small value, 64 or 256, which is much smaller than the one of BOW.

Each local descriptor x_i of an image is assigned to its nearest visual word in the codebook

$$q(x_i) = \arg\min ||x_i - \mu_j||^2, \quad \mu_j \in C. \tag{1}$$

VLAD records the sum of differences between the local descriptors and the corresponding visual words, rather than the number of SIFT descriptors assigned to the clusters. For each visual word μ_i, the difference between the SIFT descriptor x_i assigned to the visual word μ_i and its centroid is computed, and a 128-dimensional vector v_i is used to denote the sum of the residual vectors

$$v_i = \sum_{x_j : q(x_j) = \mu_i} x_j - \mu_i. \tag{2}$$

The VLAD is the concatenation of the residual vectors. The K 128-dimensional sums of residual vectors are concatenated into a single $K \times 128$ dimensional descriptor, as an image representation $V = [v_i, v_2, \ldots, v_K]$.

Since VLAD is a simplified nonprobabilistic Fisher kernel, VLAD can be power and L2-normalized. To improve the retrieval performance, VLAD needs two normalization steps. First, power normalization is applied. Each component is normalized as

$$v_j := sign(v_j)|v_j|^{\alpha}, \quad j \in [1, 2, \ldots, K \times 128]. \tag{3}$$

Here, α is usually set as 0.5 in [7]. The power normalization step is used to release the burstiness problem of VLAD vector [33]. Then, L2 normalization is employed for the entire VLAD vector

$$V := \frac{V}{||V||_2}. \tag{4}$$

If the similarity measure employs inner product, the L2 normalization can guarantee that if we choose an image from a database as the query image, the first result will be the image itself.

B. Holons Visual Representation

To extract the statistics of the local descriptors of the image database, we follow the method of BOW and VLAD and employ Kmeans to train a codebook. We cluster the independent SIFT descriptors of database by Kmeans, and the number of the obtained clusters K is set beforehand. The SIFT descriptors extracted from a large-scale image database often reach to billions, when the number of images is beyond millions. If we set K as a large value like BOW, the process of clustering will be very time-consuming and the dimension of image representation will be extremely large. Thus, the retrieval performance in terms of efficiency and memory cost will be affected. In contrast, when we set K as a small value, the small number of visual words is inadequate to describe the characteristics and patterns of billions of SIFT descriptors. To solve this problem, we adopt a two-layer hierarchical scheme inspired by [6], [40], [42], and employ hierarchical Kmeans to cluster the SIFT descriptor sets.

We first employ Kmeans to roughly partition the whole SIFT descriptor set into K clusters $C = \{c_1, c_2, \ldots, c_K\}$. Kmeans algorithm works by minimizing the intra-cluster sum of squares

$$\arg\min \sum_{i=1}^{K} \sum_{d_m \in c_i} \|d_m - \mu_i\|^2, \quad i = 1, 2, \ldots, K. \tag{5}$$

Here, μ_i is the centroid of the ith cluster c_i. In this paper, the number K of clusters obtained by Kmeans is set as a value of the power of 2.

Then, for the SIFT descriptors in each cluster, we continue employing Kmeans. Each cluster c_i is partitioned into N refined clusters, $c_i = \{c_{i,1}, c_{i,2}, \ldots, c_{i,N}\}$. Compared with traditional Kmeans clustering method, the two-layer clustering scheme can quickly obtain $K \times N$ clusters from a large SIFT descriptor sets by the two-layer hierarchical Kmeans. Similar to BOW and VLAD, we use the cluster centroids $\{\mu_1, \mu_2, \ldots, \mu_K\}$ to serve as the raw visual words and denote the correlative raw clusters $C = \{c_1, c_2, \ldots, c_K\}$. The centroids of refined clusters $\{\mu_{i,1}, \mu_{i,2}, \ldots, \mu_{i,N}\}$ are used as the refined visual words and to represent the corresponding refined clusters $c_i = \{c_{i,1}, c_{i,2}, \ldots, c_{i,N}\}$.

To qualitatively describe the distribution of SIFT descriptors with respect to the whole SIFT descriptor set, which are extracted from an independent image database, we utilize the relatively simple histogram form. Therefore, we can avoid the problem that residual vectors of different SIFT descriptors assigned into the same clusters offset each other. In this paper, two methods are described to obtain HVR. During the process of generating HVR, the first method named DHC considers the distances between local descriptors and cluster centroids. The second one called DOHC considers the distances, as well as the spatial consistency between the local features and cluster centroids.

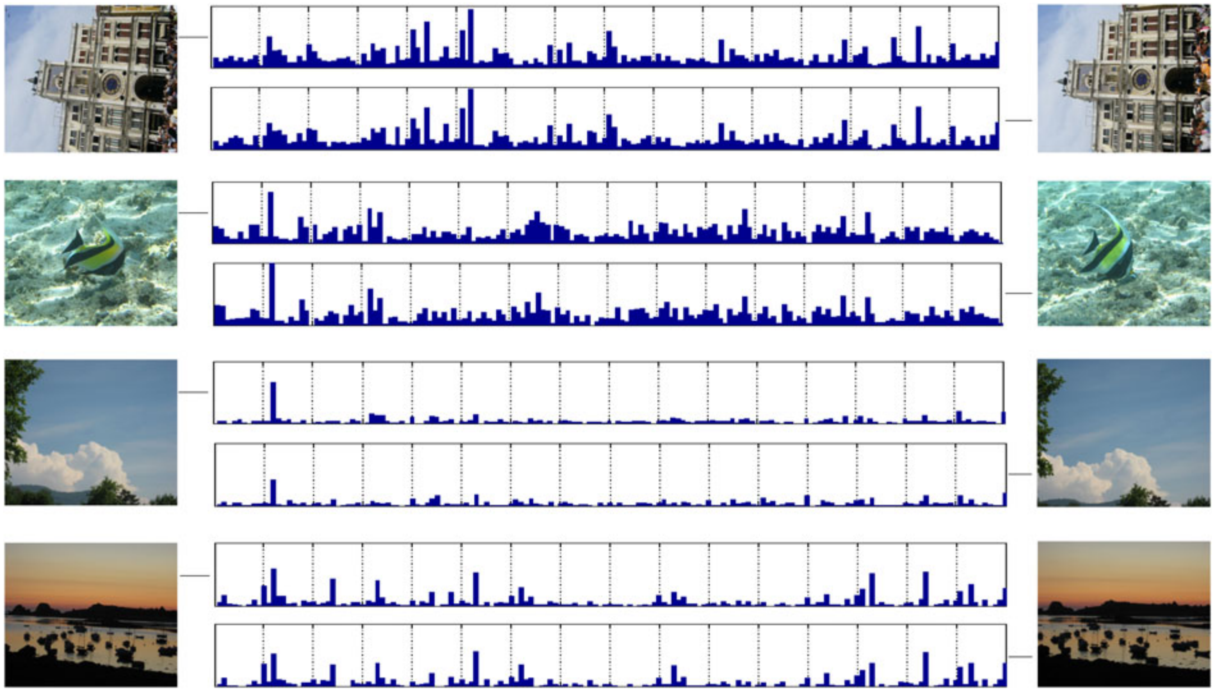


Fig. 3. Images and the corresponding DHC representations for the number of the raw partitions $K = 16$ and the number of the refined partitions $N = 10$.

1) DHC: We regard each cluster c_i as a raw partition of the whole local descriptor space, and each refined cluster $c_{i,j}$ is a refined partition of the ith raw partition c_i. We obtain $K \times N$ refined clusters, i.e., $K \times N$ refined partitions, by employing the two-layer hierarchical Kmeans method. Let $I = [d_1, d_2, \ldots, d_M] \in R^{D \times M}$ denote the SIFT descriptor extracted from an image. The raw and refined partitions and their centroids of the whole local feature space describe the global characteristics of the image datasets. According to the distance between the SIFT descriptor d_m and the centroid μ_i of the raw partition c_i, we assign the SIFT descriptors into the nearest raw partition

$$q_c(d_m) = \arg\min \|d_m - \mu_i\|^2, \quad m = 1, 2, \ldots, M. \tag{6}$$

Here, $q_c(d_m)$ represents the nearest raw partition of d_m, and we use $c_{\min_c}$ to denote it for simplicity. After obtaining the nearest raw partition $c_{\min_c}$, we calculate the distance between the SIFT descriptor d_m and the centroid $\mu_{\min_c, j}$ of a refined partition $c_{\min_c, j}$ which belongs to the raw partition $c_{\min_c}$, and assign d_m to the nearest refined partition $q_d(d_m, c_{\min_c})$

$$q_d(d_m, c_{\min_c}) = \arg\min \|d_m - \mu_{q_c(d_m), j}\|^2, \quad d_m \in c_{\min_c} \tag{7}$$

where $j = 1, 2, \ldots, N$. The first representation generation method DHC only utilizes the distance of the SIFT descriptor d_m to the centroids. Thus we compute the number of SIFT descriptors assigned into the refined partition $c_{i,j}$ as the value of the $(i\text{th}, j\text{th})$ bin

$$h(i,j) = \sum_{m=1}^{M} \begin{cases} 1, & \text{if} \quad d_m \in c_{i,j} \\ 0, & \text{otherwise.} \end{cases} \tag{8}$$

When DHC is employed as image representation, an image is described as a histogram of $K \times N$ dimensions $I = [h_{1,1}, h_{1,2}, \ldots, h_{i,j}, \ldots, h_{K,N}]$.

Fig. 3 depicts the 160-dimensional DHC representations associated with some image samples in Holiday dataset. The number of the raw partitions K is set as 16, while the number of the refined partitions N is set as 10. The black dotted lines delimit blocks of DHC representation associated with each raw partition. From Fig. 3, we can observe that the dimensions of DHC representation of the images which belong to the same categories, i.e., left and right images in the same raw, have the approximate values. It demonstrates the discriminative ability of DHC. In Fig. 3, it is also obvious that several dimensions of DHC representation are with larger values, compared with the other ones. The dimensions with larger values will affect the measure of similarity, since the dimensions have the unequal contribution for the similarity between two images. Intuitively, as shown in following section, a normalization process is necessary to release the burstiness problem.

2) DOHC: A SIFT feature $f_i = \{d_i, u_i, s_i, \theta_i\}$ includes not only the descriptor d_i, but also the location u_i, the characteristic scale s_i and the main orientation θ_i. The descriptor of SIFT feature describes the relationship of neighborhood pixels, and the main orientation θ_i describes the orientation of the descriptor, i.e., where the descriptor encoding starts. When the main orientations of SIFT features are different, the corresponding descriptors are also varying. As aforementioned, DHC describes the distribution of the SIFT descriptors according to the distances between the descriptors and cluster centroids, while neglects their main orientations, which can serve as another valuable element in accurate retrieval. According to the analysis above, we develop DOHC to generate the holons histogram, DOHC, which considers both the distance and the main orientation consistency between the local descriptors and centroids.

We compute the statistics of main orientations of the extracted SIFT features from an independent image database. According to the statistic information of the main orientations of SIFT features, the main orientation space is divided into R regions, and each region contains the approximate number of SIFT features. Then, each refined partition $c_{i,j}$ is further partitioned into R regions $c(i,j,r)$. Thus, the whole feature space is composed of $K \times N \times R$ regions. The dimension of DOHC is $K \times N \times R$. Based on the main orientation, the SIFT features in the refined partition $c_{i,j}$ are assigned to the nearest regions. Similar with the method of DHC, we calculate the number of SIFT descriptors assigned into the region $c(i,j,r)$ to serve as the value of the $(i\text{th}, j\text{th}, r\text{th})$ bin of DOHC

$$h(i,j,k) = \sum_{m=1}^{M} \begin{cases} 1, & \text{if} \quad d_m \in c_{i,j,k} \\ 0, & \text{otherwise.} \end{cases} \tag{9}$$

When DOHC is employed to represent images, each image is described as a compact histogram of $K \times N \times R$ dimensions $I = [h_{1,1,1}, h_{1,1,2}, \ldots, h_{K,N,R}]$.

C. Optimized Raw Partition Normalization Scheme

For a histogram vector, a few large dimensions may strongly affect the similarity score of histograms as shown in Fig. 3. In that case, the contribution of other dimensions will be largely reduced. To address the problem of burstiness, we propose a new normalization scheme based on the performance of HVR, namely the optimized raw partition normalization. Our normalization scheme first SSR on HVR, because it can reduce the burstiness effect by discounting large values with element-wise square rooting on the HVR histogram, and then the sub-normalization is employed on the HVR, where the histogram vector is L2 normalized on each sub-vector of the raw partition. Take DHC representation as an example, in this paper, the optimized raw partition normalization is done as follows:

$$h(i,j) = |h_{i,j}|^{\alpha} \tag{10}$$

$$V_i = \frac{V_i}{||V_i||_2} \tag{11}$$

where α is a parameter and its value is set as $0 \leq \alpha \leq 1$. V_i represents the sub-histogram belonging to the ith raw partition, i.e., $V_i = [h_{i,1}, h_{i,2}, \ldots, h_{i,N}]$. Our normalization method focuses on each sub-vector of the raw partition, rather than the whole HVR vector. This method can greatly suppress bursts and obtain higher retrieval accuracy.

Fig. 4 displays the comparison of different normalization methods for DHC on the Holidays dataset. For the original DHC representation from the Fig. 4(a), we discover that a few dimensions own most of the energy. The dimensions with the standard deviation of high values strongly influence the similarity scores of DHC representation. Compared with the normalization methods in [7], [8] displayed in Fig. 4(b) and (c), the result of Intra-normalization, our normalization results in Fig. 4(d), shows no peaks in the energy spectrum. The optimized raw partition normalization scheme we proposed is straightforward, and effectively suppresses the burstiness problem.

D. Retrieval Based on HVR

In order to improve the efficiency of image retrieval and reduce the storage cost, it is necessary to encode a high-dimensional image representation into a code of B bits. Since the DHC and DOHC representations aggregate local features in each refined partition and keep the structural property, we use the Approximate Nearest Neighbor Search method in [9] to encode the DHC and DOHC representations. The algorithm regards similarity search as a technique derived from source coding. Through the Approximate Nearest Neighbor Search algorithm, the encoded descriptors can be approximately reconstructed.

For a DHC or DOHC vector x, it is uniformly split into m sub-vectors $x^1, x^2, \ldots, x^m$. Since DHC and DOHC representations are based on the raw partitions, and each raw partition is independent, the number of sub-vector m is set as the same as that of raw partitions K. Each sub-vector is the sub-histogram in the raw partition, thus it can keep the structure information.

A product quantizer $q(x)$ is composed of m distinct sub-quantizers with respect to m sub-vectors. The vector x is encoded by the product quantizer

$$q(x) = (q_1(x^1), q_2(x^2), \ldots, q_m(x^m)). \tag{12}$$

The vector x is mapped to a tuple of indices by quantizing the sub-vectors, respectively. All distinct sub-quantizers are with the same number k_s of reproduction values obtained by Kmeans. To reduce the assignment complexity $O(m \times k_s)$, k_s is usually set as a small number. The total number of centroids produced by the product quantizer is $(k_s)^m$, which is a very large value. Finally, a vector is encoded as a code of $B = m \log_2 k_s$ bits.

The IVFADC [9] is used to build the index files and search for nearest neighbors. The query x is not encoded, thus there is no approximation error for query vector. According to decomposition, the square distance between the query vector x and database vector y are computed as follows:

$$d(x,y) = ||x - q(y)||^2 = \sum_{j=1,\ldots,m} ||x^j - q_j(y_j)||^2 \tag{13}$$

where y_j denotes the jth sub-vector of the database vector y. The square distance of each sub-vector x^j to the k_s centroids of the corresponding sub-quantizer q^j is computed prior to search, and stored in a look-up table. Thus, the square distances in Eq. (13) can be obtained by looking up the tables. The complexity of the generation of the look-up tables is $O(K \times N \times k_s)$ for DHC representation, and $O(K \times N \times R \times k_s)$ for DOHC representation. Since k_s is much smaller than the number n of images in the database, the complexity can be minor, compared with the cost $O(K \times N \times n)$ or $O(K \times N \times R \times n)$ resulting from the direct computation of the query vector and all the database vectors.

IV. Experiment

The experiments are conducted by submitting each query image to the image retrieval system, and obtains a list of images

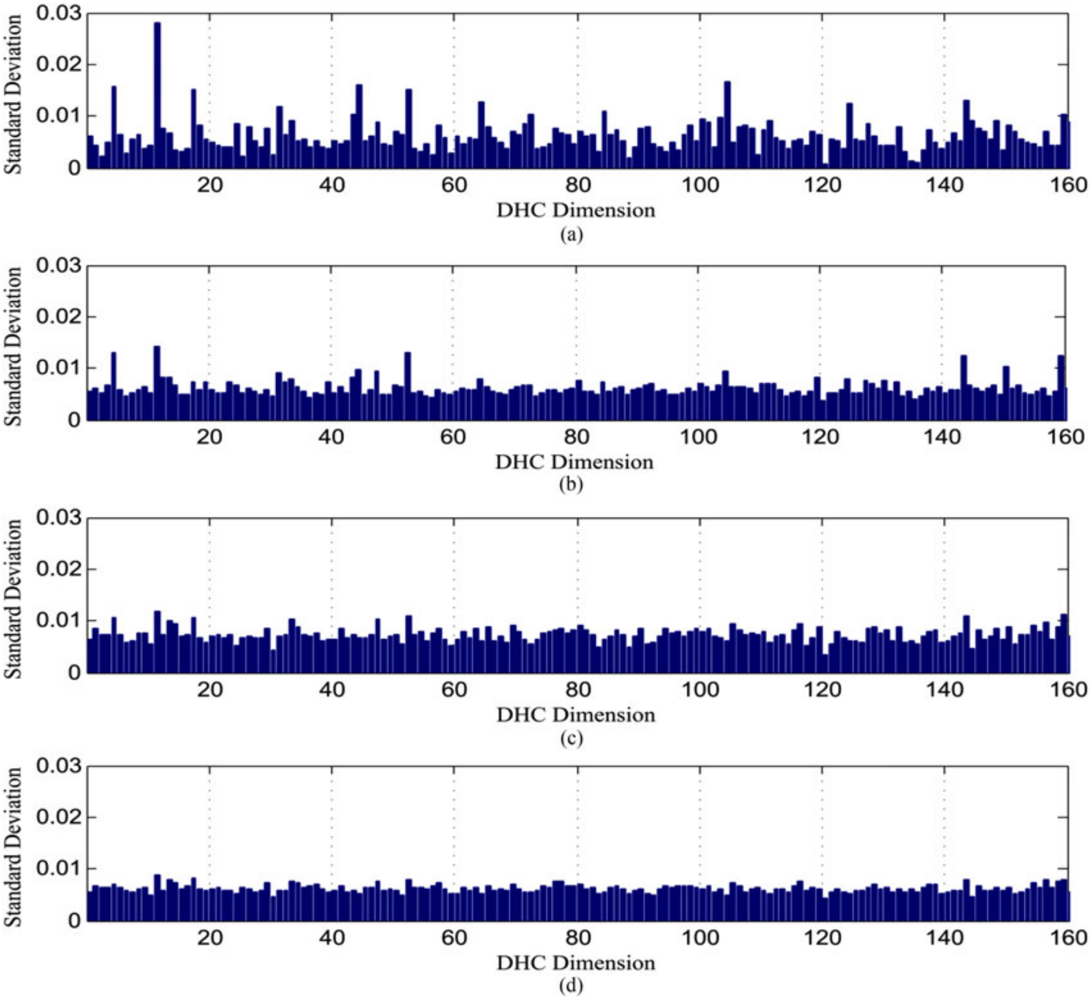


Fig. 4. Comparison of the normalization methods for DHC on the Holidays dataset. (a) denotes the standard deviation distribution of original DHC on the Holiday datatset. (b), (c), and (d) represent the distribution about standard deviation of DHC with the normalization methods in [7], [8], and the proposed normalization method, respectively

sorted by decreasing similarity. In this section, we first introduce the data sets and evaluation protocol and then give a description of implementation details in the experiments. Finally, a detailed analysis of the results is made.

A. Data Sets And Evaluation Protocol

In this paper, we use four challenging image datasets, namely, UKbench, Holidays, Oxford5k Building and Paris6k, as the testbed. We also employ two large image datasets, ImageNet-V and ImageNet-T, for all the learning stages and large-scale retrieval experiments.

UKbench [2] contains 10 200 images belonging to 2550 objects. Each object has four images taken from different viewpoints. In our experiments, all the 10 200 images are used as both queries and database images. The expected result of using each image query is the four most similar images of the same object. The most common evaluation manner of UKbench dataset is to count the average number of relevant images (including the query itself) that are ranked in the first four positions. This corresponds to N-S score and the mean average precision (MAP).

Holidays [11] is collected by INRIA. It contains 1491 personal holiday images undergoing various transformations. In Holidays dataset, only 500 images are used as queries and the other 991 are the corresponding similar images. The retrieval performance of the dataset is usually estimated by MAP.

Oxford5k Building [5] is a collection of 5062 building images, which are downloaded from Flickr. For simplification, we use Oxford5k to denote it. The dataset defines 55 queries by a rectangular region of interest. The query images correspond to 11 distinct buildings in Oxford, and the images in the dataset are annotated as good, OK, bad, or junk. The good image is a clear picture of the building, and in the OK images, more than 25% of the building is visible. In contrast, the junk images are with very high levels of occlusion or distortion, which indicates it is unclear whether a user would consider the image as relevant or not. MAP is used to evaluate performance of the dataset.

Paris6k [28] is composed of 6412 images, which are all collected from Flickr by searching for the particular Paris landmarks. Similar with Oxford5k Building dataset, it also contains 55 query images annotated by a rectangular of interest. The ground truths of images are offered as Oxford5k Building. The retrieval performance is evaluated by MAP.

ImageNet is an image database organized according to the WordNet hierarchy [10] in which each node of the hierarchy is depicted by hundreds and thousands of images. Since the ImageNet dataset is publicly available and contains sufficient large variations, it is well suited to benchmark the retrieval accuracy, computation, and memory usage for the large-scale image retrieval. We use 1.26 million images of 1000 categories as a large-scale image dataset, denoted by ImageNet-T. The large-scale image retrieval experiment is conducted by merging the images in ImageNet-T with the other data sets. In addition, we employ 100K images, denoted by ImageNet-V, to sample the SIFT features, split the feature space, generate VLAD, DHC and DOHC representations, and train all quantizers. The ImageNet-V has no overlap with the four aforementioned datasets and ImageNet-T.

B. Implementation Details

For each image, we describe it by a VLAD, DHC or DOHC representation. From the description in section IV, it can be noticed that the critical parameters of our HVR is K and N. In the experiments, we set the number of the raw partitions as the power of 2, namely 8, 16, 32, 64 and 128. And relevantly, the number of refined partitions is set as 10, 20, 30, 40. Thus, each VLAD is a $K \times 128$ vector and the dimension of the DHC is $K \times N$. In order to further enhance the discrimination power of HVR, we take the main orientation of SIFT features into account in DOHC representation. The main orientation θ is split into four regions, i.e., $[-\pi, -1.67]$, $[-1.67, 0.213,]$, $[0.213, 1.68,]$, $[1.68, \pi]$, and the DOHC presentation has the dimension of $K \times N \times 4$. Especially, we use DHC-N to represent the DHC representation with N refined partitions. DHC-10, DHC-20, DHC-30 and DHC-40 denote that the corresponding DHC representations with $N =$ 10, 20, 30 and 40, respectively. Correspondingly, DOHC representation adopts the same way to distinguish different number of refined partitions.

For large-scale image retrieval experiments, we train two different kinds of complete quantizers, as well as two product quantizers, with the equal number of centroids for VLAD and DOHC representations, by employing Kmeans on the independent ImageNet-V database. In this paper, k_s is set as 8. All VLAD and DOHC representation vectors in the database are first quantized into their corresponding complete quantizers, and then quantized into the nearest distinct sub-quantizers according to the residual vectors of the representation vectors to the centroids of their complete quantizers. When an image representation is quantized into the complete quantizer, the relevant inverted file will add an item, containing an image ID and a vector code. Since all representation vectors are partitioned into K sub-vectors, a VLAD or DOHC representation is converted to a code of $K \times \log_2 k_s$. Therefore, the DOHC and VLAD cost the same memory in the inverted files.

Since DOHC and VLAD cost the same memory in the inverted files, which is explained in experimental details. In addition, the index structure and search method employed by HVR and VLAD are the same, the difference of retrieval efficiency mainly derives from the distribution of image representations from the whole datasets with respect to the quantizers. It shows in the diversity of items in the inverted lists associated with the centroids of the quantizers, which is related to the corresponding trained quantizers. However, the training process of quantizers has a certain randomness because K-means algorithm adopts random method starts with a random initialization. Therefore, the retrieval time cannot effectively prove which method has better efficiency. Thus, in this paper, we primarily focus on the comparison of accuracy.

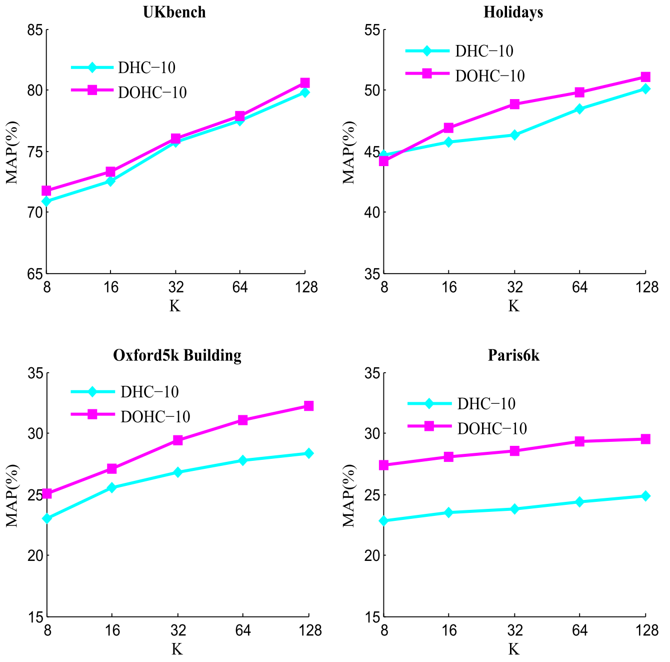


Fig. 5. MAPs of DHC and DOHC on different datasets with the increasing number of raw partitions and the value parameter $N = 10$.

C. Experimental Results

In the following section, we make a detailed analysis of the performance of the proposed representation HVR. First, performance of DHC and DOHC with different number of raw and refined partitions is evaluated. Then, a comparison is made between the performance of DHC, DOHC and VLAD on the medium-scale image databases. Finally, we report the retrieval performance of DOHC and VLAD on large-scale retrieval experiments merged by the ImageNet-T with Holidays.

1) Impact of Parameters for DHC and DOHC: We conduct experiments on the four aforementioned datasets, namely, UKbench, Holidays, Oxford5k Building and Parisy6k, to evaluate the influence of parameters on the retrieval performance. In order to explore the relationship of the number of raw and refined partitions with retrieval accuracy, we control the value of one parameter. Here, we set the number of refined partitions as 10 and obtain the retrieval accuracies of DHC and DOHC representation with the increasing of raw partitions, as shown in Fig. 5. Intuitively, we can observe that, both the MAPs of DHC-10 and

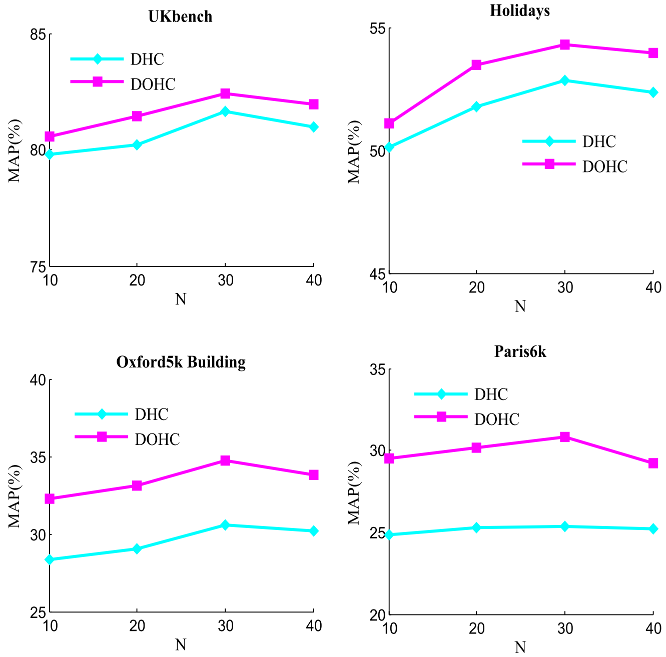


Fig. 6. MAPs of DHC and DOHC on different datasets with the increasing number of refined partitions and the value of parameter $K = 128$.

DOHC-10 representation increase with the continually increasing number of the raw partitions. Take the experimental results on UKbench as an example, when the number of raw partitions K increases from 8 to 128, the MAP of DHC-10 representation improves from 70.90% to 79.77%, while that of DOHC-10 representation improves from 71.70% to 80.57%. Moreover, the improvement of MAPs with the increase of parameter K, which is obtained by DHC-10 and DOHC-10 representation on the rest of the three datasets, are also evidently demonstrated in the Fig. 5. It can be stated the larger number of the raw partitions, the higher descriptive and discriminative ability DHC and DOHC have. The reason is that the whole SIFT feature space is split up more precisely when the number of raw partitions increases. Thus, the SIFT descriptors of an image can be assigned into more approximate raw partitions, and the distribution with respect to the whole SIFT feature space is more accurate.

Meanwhile, we take the same way, namely, give the parameter K a certain value, to explore the influence of parameter N on the retrieval performance of the DHC and DOHC representation. Particularly, we fix the number of the raw partitions K to 128, which achieves the best performance among all values. From the Fig. 6, it can be observed that both DHC and DOHC with the number of refined partitions $N = 30$ obtain the higher retrieval accuracy than other partitions, while the MAP with the least refined partitions is lowest. The experimental results on the four datasets all follow this rule. For example, on UKbench dataset, DHC-30 achieves the highest MAP of 81.63% among all the DHC representations. Compared with DHC-10, the MAP of DHC-30 improves approximate 2%. For DHC-20, DHC-40, their MAP is also 1.43% and 0.66% lower than that achieved by DHC-30, respectively. And the changing rule of the MAP of DOHC has the similar trend with that of DHC. In addition, DHC-30 achieves the best accuracies of 52.84%, 30.55% and 25.32% on the Holidays, Oxford5k Building and Paris6k, respectively. Thus, it worths to note that the MAP of DHC and DOHC does not always monotonously increase with the increase of the number N of refined partitions.

TABLE I
MAP OF DHC, DOHC, BOW, AND VLAD (MAP IS THE AVERAGE OF ACCURACY PRECISION %)

Datasets	Ukbench	Holidays	Oxford5k	Paris6k
BOW	79.62	46.91	33.85	25.42
VLAD	**84.68**	**55.75**	32.10	29.12
DHC-10	79.77	50.11	28.36	24.86
DHC-20	80.20	51.76	29.04	25.26
DHC-30	81.63	52.84	30.55	25.32
DHC-40	80.97	52.34	30.17	25.23
DDHC-10	80.57	51.10	32.28	29.53
DOHC-20	81.43	53.45	33.09	30.17
DOHC-30	82.41	54.32	**34.76**	**30.78**
DOHC-40	81.93	53.94	33.84	29.23

2) Comparison With the State-of-the-Art: We compare with the state-of-the-art approaches, e.g., BOW and VLAD. Specifically, the codebook size of BOW is set to 20 000. Table I demonstrates that both HVR outperforms BOW in terms of both memory usage and retrieval performance and VLAD outperforms BOW as verified in [17]. Moreover, HVR and VLAD both employ the idea of aggregation. Thus, in this section, we focus on comparing the retrieval performance of DHC and DOHC, with the VLAD on the medium-scale image datasets.

Figs. 5 and 6 clearly demonstrate that the DOHC representation obtains higher retrieval accuracy than DHC. For example, DOHC-30 representation achieves the MAP of 82.41% on the Ukbench, which is 0.78% higher than that of DHC with the same parameters. This advantage of DOHC representation is more obvious on Holidays and Oxford5k datasets. Take the Holidays as another example, the DOHC-30 obtains the best MAP of 30.78%, which is 5.46% higher than the best MAP of the DHC representation. In fact, the advantages of MAP achieved by the DOHC representation are beneficial from the generation of compact histograms. When the whole feature space is spilt, the DOHC representation not only refers to the distances between SIFT descriptors and centroids of the partitions, but also considers the main orientations of SIFT, which describes the spatial information of SIFT descriptors, i.e., the orientations where the descriptors start encoding. The main orientations of SIFT features serve as a valuable supplement for the descriptors.

Since both VLAD and DOHC representations with the number of raw partitions $K = 128$ have achieved the best retrieval MAP, We fix the parameter K to 128. According to Table I, we can discover that, compared with VLAD, the proposed DOHC representation achieves the approximate retrieval performance on Ukbench and Holidays datasets. Particularly, the best MAP of DOHC representation is 82.41% on the Ukbench dataset, which is close to that of VLAD. On the Holidays, the DOHC representation also achieves approximate MAP of 54.12% to that of 55.75%, which is obtained by VLAD. Table I also

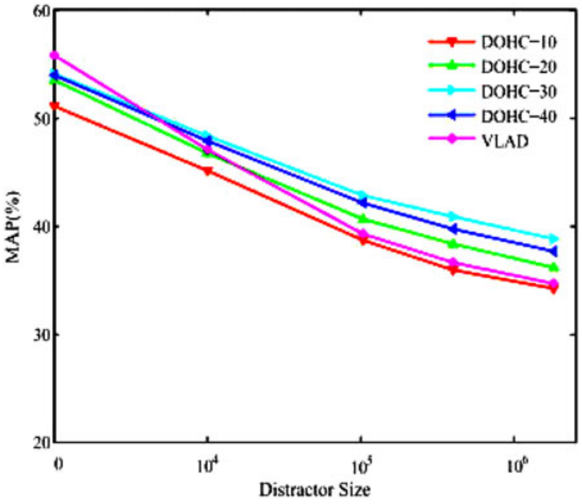


Fig. 7. Large-scale image retrieval performance of VLAD and DOHC with different numbers of distractor images.

clearly demonstrates that the proposed DOHC representation outperforms VLAD on Oxford5k and Paris6k. The MAP of the DOHC-30 representation reaches to 34.76%, which is 2.66% higher than that of VLAD on the Oxford5k dataset, and the DOHC representation also improves the MAP to 30.78% on the Paris6k dataset. Based on these analysis, we conclude that DOHC representation is suitable for the images with similar intra structures. In the Oxford5k and Paris6k datasets, the building images are usually with more similar intra structures, *e.g.*, the similar windows and roofs, than those in the Ukbench and Holidays datasets. In this case, DOHC representation generated by assigning the SIFT features into the corresponding refined partitions can successfully describe the distribution of SIFT features with respect to the centroids of the partitions.

3) Large-Scale Image Retrieval Experiments: To further evaluate the performance of HVR on the large-scale image datasets, we merge Holidays with ImageNet-T dataset, which are used as distractors. We use different numbers of distractors, including 0, 10 000, 120 000, 600 000, and the whole ImageNet-T dataset, to test the scalability of the proposed representation. Since it is certified that the performance of DOHC precedes that of DHC in the previous experiments, we only compare the performance of DOHC and VLAD in the large-scale retrieval experiments. We also set the parameter K to 128 when conducting large-scale retrieval experiments. Fig. 7 depicts the large-scale image retrieval accuracies of VLAD and DOHC with different numbers of distractor images.

Table II displays that the performance of both VLAD and DOHC representations gradually decreases with the augment of distractor images in the database. When the number of distractor images is 0, all DOHC and VLAD obtain their best accuracies. When the entire ImageNet-T dataset is added into the Holidays database, which means 1 260 000 distractor images, the MAP values of VLAD and DOHC reduce significantly. Especially, the performance of VLAD reduces from 55.75% to 34.63% and DOHC-30 and DOHC-40 almost reduce 15% and 16%, respectively. The decline of MAP mainly results from the inverted files containing more items with the increase of the distractors. The query vector has to compute the similarity scores with more candidate images. Since the similar images of query image does not increase, the difficulty of searching the similar images from larger candidate image sets enlarges, and then the MAP of image retrieval decreases.

TABLE II
COMPARISON OF MAP OBTAINED BY DOHC AND VLAD WITH DIFFERENT NUMBERS OF DISTRACTOR IMAGES

		DOHC			
Distractor Size	VLAD	$N = 10$	$N = 20$	$N = 30$	$N = 40$
0	55.75	51.10	53.45	54.12	53.94
10 k	47.03	45.12	46.72	48.34	47.88
120 k	39.26	38.64	40.66	42.86	42.13
600 k	36.57	35.96	38.32	40.85	39.69
1.26 M	34.63	34.18	36.15	38.78	37.64

Although the MAPs of both VLAD and DOHC decrease with the increase of distractor images, we can learn that the MAP of the DOHC representation reduces more gently than that of VLAD. When the distractor images increase from 0 to 1 260 000, on the Holidays dataset, the MAP of VLAD declines to 34.63%, which is 21.12% less than the MAP obtained by VLAD without distractors. However, the decline of the MAPs achieved by the different DOHC representations is no more than 20%. The MAP only reduces by 17.30% for the DOHC-20 representation, and the decrease of the MAPs obtained by the DOHC-10, DOHC-30, DOHC-40 representations is less than 17%. Thus, compared with VLAD, DOHC can maintain its discriminative ability more effectively when the size of image database increases. Namely, the DOHC representation outperforms VLAD in the terms of scalability.

From aforementioned analysis, we obtain the following observations. First, the DOHC representation outperforms DHC with the same parameters, and DOHC shows higher retrieval performance for the images with more similar local information. Furthermore, compared with VLAD, DOHC has better scalability, i.e., with the increase of distractor images in the database, the retrieval performance of the DOHC declines more gently than that of VLAD, as shown in Fig. 7.

V. CONCLUSION

In this paper, we introduce a new image representation method, named HVR, to improve the retrieval accuracy of the images with similar intra-structures. The proposed HVR, including DHC and DOHC, makes full use of the global characteristics and statistics of the independent local feature set. In addition, the HVR also describes the global distribution of the local features of an image with respect to the entire local features set of image datasets. HVR enhances the discriminative power of individual features, by utilizing the two-layer hierarchical partitions of the local feature space and the correlation of distance and spatial information. The comprehensive experimental results demonstrate that HVR improves the retrieval accuracy and achieves better scalability for large-scale image datasets, which demonstrates the effectiveness of our image representation.

REFERENCES

[1] J. Sivic and A. Zisserman, "Video Google: A text retrieval approach to object matching in videos," in *Proc. 9th IEEE Int. Conf. Comput. Vis.*, 2003, pp. 1470–1477.

[2] D. Nister and H. Stewenius, "Scalable recognition with a vocabulary tree," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2006, vol. 2, pp. 2161–2168.

[3] D. Lowe, "Distinctive image features from scale-Invariant keypoints," *Int. J. Comput. Vis*, vol. 60, no. 2, pp. 91–100, 2004.

[4] H. Jegou, M. Douze, and C. Schmid, "Improving bag-of-features for large scale image search," *Int. J. Comput. Vis*. vol. 8, no. 3, pp. 316–33, 2010.

[5] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Object retrieval with large vocabularies and fast spatial matching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2007, pp. 1–8.

[6] X. Wang *et al.*, "Contextual weighting for vocabulary tree based image retrieval," in *Proc. IEEE Int. Conf. Comput. Vis.*, Nov. 2011, pp. 209–216.

[7] H. Jegou *et al.*, "Aggregating local image descriptors into compact codes," *IEEE Trans. Pattern Anal. Mach. Intell*, vol. 3, no. 9, pp. 1704–1716, Sep. 2012.

[8] R. Arandjelovic and A. Zisserman, "All about VLAD," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2013, pp. 1578–1585.

[9] H. Jegou, M. Douze, and C. Schmid, "Product quantization for nearest neighbor search," *IEEE Trans. Pattern Anal. Mach. Intell*, vol. 33, no. 1, pp. 117–128, Jan. 2011.

[10] J. Deng *et al.*, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2009, pp. 248–255.

[11] H. Jegou, M. Douze, and C. Schmid, "Hamming embedding and weak geometric consistency for large scale image search," in *Proc. 10th Eur. Conf. Comput. Vis.*, 2008, pp. 304–317.

[12] Y. Jiang, J. Meng, and J. Yuan, "Randomized visual phrases for object search," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2012, pp. 3100–3107.

[13] Y. Zhang, Z. Jia, and T. Chen, "Image retrieval with geometry preserving visual phrases," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2011, pp. 809–816.

[14] L. Wu, Y. Hu, M. J. Li, N. H. Yu, and X. S. Hua, "Scale-invariant visual language modeling for object categorization," *IEEE Trans. Multimedia*, vol. 11, no. 2, pp. 286–297, Feb. 2009.

[15] S. Zhang *et al.*, "Building contextual visual vocabulary for large-scale Image applications," in *Proc. 18th ACM Int. Conf. Multimedia*, 2010, pp. 501–510.

[16] Y. X. Li *et al.*, "Difficulty guided image retrieval using linear multiple feature embedding," *IEEE Trans. Multimedia*, vol. 14, no. 6, pp. 1618–1630, Dec. 2010.

[17] Y. Cao, C. Wang, Z. Li, L. Q. Zhang, and L. Zhang, "Spatial bag-of-features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2010, pp. 3352–3359.

[18] G. Tolias and Y. Avrithis, "Speeded-up, relaxed spatial matching," in *Proc. IEEE Int. Conf. Comput. Vis.*, Nov. 2011, pp. 1653–1660.

[19] N. Hoang, V. Gouet-Brunet, M. Rukoz, and M. Manouvrier, "Embedding spatial information into image content description for scene retrieval," *Pattern Recog.*, vol. 43, no. 9, pp. 301–302, 2010.

[20] J. Deng, A. C. Berg, and L. Fei-Fei, "Hierarchical semantic indexing for large scale image retrieval," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2011, pp. 785–792.

[21] F. Yu, R. Ji, M. Tsai, G. Ye, and S. Chang, "Weak attributes for large scale image retrieval," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2012, pp. 2949–2956.

[22] Y. G. Jiang, C.W. Ngo, and J, Yang, "Towards optimal bag-of-features for object categorization and semantic video retrieval," in *Proc. CIVR*, 2007, pp. 494–501.

[23] Z. Lin and J. Brandt, "A local bag-of-features model for large scale object retrieval," in *Proc. 11th Eur. Conf. Comput. Vis.*, 2010, pp. 294–308.

[24] R. Arandjelovic and A Zisserman, "Three things everyone should know to improve object retrieval," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2012, pp. 2911–2918.

[25] J. Delhumeau, P. H. Gosselin, H. Jegou, and P. Prez, "Revisiting the VLAD image representation," in *Proc. 21st ACM Int. Conf. Multimedia*, 2013, pp. 653–656.

[26] F. Perronnin and C. R. Dance, "Fisher kernels on visual vocabularies for image categorization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2007, pp. 1–8.

[27] F. Perronnin, Y. Liu, J. Sanchez, and H. Poirier, "Large-scale image retrieval with compressed Fisher vectors," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2010, pp. 3384–3391.

[28] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Lost in quantization: Improving particular object retrieval in large scale image databases," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2008, pp. 1–8.

[29] J. C. van Gemert, C. J. Veenman, A. W. M. Smeulders, and J. M. Geusebroek, "Visual word ambiguity," *IEEE Trans. Pattern Anal. Mach. Intell*, vol. 32, no. 7, pp. 1271–1283, Jul. 2010.

[30] Y. Su and F. Jurie, "Visual word disambiguation by semantic contexts," in *Proc. IEEE Int. Conf. Comput. Vis.*, Nov. 2011, pp. 311–318.

[31] 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.

[32] J. Philbin, M. Isard, J. Sivic, and A. Zisserman, "Descriptor learning for efficient retrieval," in *Proc. 10th Eur. Conf. Comput. Vis.*, 2010, pp. 677–691.

[33] H. Jegou, M. Douze, and C. Schmid, "On the burstiness of visual elements," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2009, pp. 1169–1176.

[34] Q. Zhang and E. Izquierdo, "Histology image retrieval in optimised multi-feature spaces," *IEEE J. Biomed. Health Informat.*, vol. 17, no. 1, pp. 240–249, Jan. 2013.

[35] J. C. Caicedo and E. Izquierdo, "Combining low-level features for improved classification and retrieval of histology images," *Trans. Mass-Data Anal. Images Signals*, vol. 2, no. 1, pp. 68–82, 2009.

[36] Q. Zhang and E. Izquierdo, "A multi-feature optimization approach to object-based image classification," in *Proc. 5th Int. Conf. Image Video Retrieval*, 2006, pp. 310–319.

[37] Y. Yang, L. Yang, G. Wu, and S. Li, "Image relevance prediction using query-context bag-of-object retrieval model," *IEEE Trans. Multimedia*, vol. 16, no. 6, pp. 1700–1712, Oct. 2014.

[38] H. Xie, Y. Zhang, J. Tan, L. Guo, and J. Li, "Contextual query expansion for image retrieval," *IEEE Trans. Multimedia*, vol. 16, no. 4, pp. 1104–1114, Jun. 2014.

[39] Y. Liang, L. Dong, S. Xie, N. Lv, and Z. Xu, "Compact feature based clustering for large-scale image retrieval," in *Proc. IEEE Int. Conf. Multimedia Expo Workshops*, 2014, pp. 1–6.

[40] L. Dong, J. Su, and E. Izquierdo, "Scene-oriented hierarchical classification of blurry and noisy images," *IEEE Trans. Image Process.*, vol. 21, no. 5, pp. 2534–2545, May 2012.

[41] L. Dong and E. Izquierdo, "A biologically inspired system for classification of natural images," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 17, no. 5, pp. 590–603, May 2007.

[42] L. Dong and E. Izquierdo, "Global-to-local oriented perception on blurry visual information," in *Proc. IEEE Int. Conf. Image Process.*, Oct. 2008, pp. 2168–2171.

[43] R. Chang and X. Qi, "Semantic clusters based manifold ranking for image retrieval," in *Proc. 18th IEEE Int. Conf. Image Process.*, 2011, pp. 2425–2428.

[44] P. Niloufar and B. S. Manjunath, "PixNet: A localized feature representation for classification and visual search," *IEEE Trans. Multimedia*, vol. 17, no. 5, pp. 616–625, May 2015.

[45] L. Zhen, H. Q. Li, W. G. Zhou, R. C. Hong, and Q. Tian, "Uniting Keypoints: Local visual information fusion for large-scale image search," *IEEE Trans. Multimedia*, vol. 17, no. 4, pp. 538–548, Apr. 2015.

[46] Z. C. Li, J. Liu, J. H. Tang, and H. Q. Lu, "Robust structured subspace learning for data representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 10, pp. 2085–2098, Oct. 2015.

[47] Z. C. Li, J. Liu, Y. Yang, X. F. Zhou, and H. Q. Lu, "Clustering-guided sparse structural learning for unsupervised feature selection," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 9, pp. 2138–2150, Sep. 2014.

[48] Z. C. Li and J. H. Tang, "Unsupervised feature selection via nonnegative spectral analysis and redundancy control," *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 5343–5355, Dec. 2015.

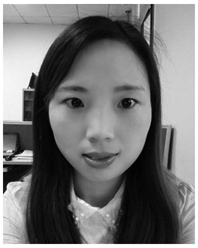
Le Dong received the Ph.D. degree in electronic engineering and computer science from the Queen Mary University of London, London, U.K.

She is currently an Associate Professor of computer science with the University of Electronic Science and Technology of China, Chengdu, China. She has authored or coauthored papers that appeared in international journals and conferences such as the IEEE Transactions on Image Processing, the IEEE Transactions on Circuits and Systems for Video Technology, the IEEE Transactions on Neural Networks, the IEEE Transactions on Signal Processing, ACM MultiMedia, ICME, and ICPR. Her research interests include bioinspired models, holons representation framework, and mobile social networks, and she has worked on several projects including the NSFC Surface Project, the NSFC Youth Project, and the NSFC Important Research Project.

Prof. Dong is currently the Secretary-General of VALSE, the Executive Secretary of the Next-Generation National-Local Joint Engineering Center, and the Director of International Talents Program, HEIFER. She has been a Reviewer for several international journals and conferences including the IEEE Transactions on Circuits and Systems for Video Technology, the IEEE Transactions on Neural Networks, the IEEE Transactions on Image Processing, and PR. She has Co-Chaired or Co-Steered a number of conferences and workshops.

Yan Liang received the B.S. degree in computer science and technology and the M.S. degree in computer technology from the University of Electronic Science and Technology of China, Chengdu, China, in 2012 and 2015, respectively.

She is a Research Staff Member with the Institute of Computer Application, China Academy of Engineering Physics, Mianyang, China. Her research interests included image analysis and understanding, big data analysis, machine learning, and data mining.



Gaipeng Kong is currently working toward the Postgraduate degree with the School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu, China.

Her research interests include image retrieval, particularly including image clustering, visual matching, and holon representation.

Qianni Zhang received the M.Sc. degree in internet signal processing and the Ph.D. degree in visual information retrieval from the Queen Mary University of London (QMUL), London, U.K., in 2004 and 2007, respectively.

She is a Senior Lecturer with the School of Electronic Engineering and Computer Science, QMUL. She was previously a Senior Researcher with QMUL, coordinating and driving core technical work in EU-funded projects REVERIE, 3DLife, MISSA, and RUSHES. She has authored or coauthored more than 45 peer-reviewed articles in major international conferences, journals, and books. Her research interests include semantic media understanding, feature fusion and semantic context inference for image classification, medical image analysis, machine learning for medical data analysis, forensic video applications, and 3-D human modelling.

Ms. Zhang is a Reviewer for major vision and machine learning journals (the IEEE Transaction on Circuits and Systems for Video Technology and the IEEE Transactions on Image Processing, *Elsevier SPIC*, and *MTAP*) and a PC Member or Reviewer for major conferences (ICIP, ICASSP, ACM MM, etc.).

Xiaochun Cao received the B.E. and M.E. degrees in computer science from Beihang University, Beijing, China, and the Ph.D. degree in computer science from the University of Central Florida, Orlando, FL, USA.

He has been a Professor with the Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China, since 2012. After graduation, he spent about three years with the Department of Government Research, ObjectVideo Inc., Reston, VA, USA, as a Research Scientist. From 2008 to 2012, he was a Professor with Tianjin University, Tianjin, China. He has authored and coauthored more than 80 journal and conference papers.

Prof. Cao is a Fellow of the IET. He is on the Editorial Board of the IEEE Transactions of Image Processing. His dissertation was nominated for the University of Central Florida's university-level Outstanding Dissertation Award. In 2004 and 2010, he was the recipient of the Piero Zamperoni Best Student Paper Award at the International Conference on Pattern Recognition.

Ebroul Izquierdo received the Ph.D. degree from Humboldt University of Berlin, Berlin, Germany.

He is the Chair of Multimedia and Computer Vision and Head of the Multimedia and Vision Group, School of Electronic Engineering and Computer Science, Queen Mary University of London, London, U.K. He was previously a Senior Researcher with the Heinrich-Hertz Institute for Communication Technology, Berlin, Germany, and the Department of Electronic Systems Engineering, University of Essex, Colchester, U.K.

Prof. Izquierdo is a Chartered Engineer, a Fellow Member of The Institution of Engineering and Technology (IET), and a Member of the British Machine Vision Association. He was a Past Chairman of the IET professional network on information engineering. He is a Member of the Visual Signal Processing and Communications Technical Committee of the IEEE Circuits and Systems Society and a Member of the Multimedia Signal Processing Technical Committee of the IEEE. He is or has been the Associated Editor of the IEEE Transactions on Circuits and Systems for Video Technology (from 2002 to 2010) and the IEEE Transactions on Multimedia (from 2010 to 2015). He has been a Member of the Organizing Committee of several conferences and workshops in the field of image and video processing. He has been the General Chair of the European Workshop on Image Analysis for Multimedia Interactive Services, London 2003 and Seoul 2006, the European Workshop for the Integration of Knowledge, Semantics and Content, London 2004 and 2005, the Mobile Multimedia Communications Conference MobiMedia, Algero 2006, the International Conference on Content Based Multimedia Indexing, London 2008, the IET Conference on Visual Information Engineering, Xian 2008, and the International Conference on Imaging for Crime Detection and Prevention, London 2015.