Mobile Visual Search Compression With Grassmann Manifold Embedding

Zhaobin Zhang, Li Li, *Member, IEEE*, Zhu Li, *Senior Member, IEEE*, and Houqiang Li, *Senior Member, IEEE*

***Abstract*— With the increasing popularity of mobile phones and tablets, the explosive growth of query-by-capture applications calls for a compact representation of the query image feature. Compact descriptors for visual search (CDVS) is a recently released standard from the ISO/IEC moving pictures experts group, which achieves state-of-the-art performance in the context of image retrieval applications. However, they did not consider the matching characteristics in local space in a large-scale database, which might deteriorate the performance. In this paper, we propose a more compact representation with scale invariant feature transform (SIFT) descriptors for the visual query based on Grassmann manifold. Due to the drastic variations in image content, it is not sufficient to capture all the information using a single transform. To achieve more efficient representations, a SIFT manifold partition tree (SMPT) is initially constructed to divide the large dataset into small groups at multiple scales, which aims at capturing more discriminative information. Grassmann manifold is then applied to prune the SMPT and search for the most distinctive transforms. The experimental results demonstrate that the proposed framework achieves state-of-the-art performance on the standard benchmark CDVS dataset.**

***Index Terms*— Mobile visual search, Grassmann manifold, subspaces embedding, CDVS, SIFT, compact descriptors.**

I. Introduction

THERE are millions of images and videos added to the servers daily. For example, every second 777 photos are posted on Instagram [1] and Snapchat now achieved over 10 billion video views per day during the past year [2]. All these benefit from the rapid development of the technology of mobile devices [3]. Hand-held mobile devices, such as camera-phone, PADs are expected to become ubiquitous platforms for visual search and mobile augmented reality applications [4]–[6]. They have evolved into powerful image and video processing devices equipped with high-resolution cameras, color displays, hardware-accelerated graphics, Global Position System (GPS) and connected to broadband wireless networks [7]. All these functionalities enable a new class of applications which use the camera phone to initiate search queries about objects in visual proximity to users [8].

Mobile Visual Search (MVS) can be used for identifying interesting products, landmark search, comparison shopping, searching information about movies, CDs, shops, real estate, print media or artworks [9]. First commercial deployments of such systems include Google Goggles [10], Ricoh iCandy [11], Amazon Snaptell [12] and Layar [13]. Recently, Pinterest [14] also moves to leverage visual search technologies in order to connect consumers in the e-commerce business.

In traditional text-based search, users can get the accurate retrieval results once the exact words are given. However, it is far more difficult to describe an image if people want to search for some relevant information. In the past quite a long time, people have been working on extracting image features and describing the image compactly and accurately [15]–[17]. Different from text-based content which is very simple and concise, the image contains much more information and is much more challenging to generate concise representations. Therefore, to fully characterize the information of an image, the existing algorithms tend to generate high-dimensional features which are usually oversized to transmitted over the limited-bandwidth wireless networks. Meanwhile, the requirements for mobile visual search (MVS) such as lower latency, better user experience, higher accuracy pose a unique set of challenges in practical applications. Therefore, an applicable strategy is feature extraction [18] and feature compression are performed at the client end while matching and retrieval is carried out on the server.

There are two aspects researchers are working on, i.e., generating compact feature descriptors and compressing the feature descriptors. Developing compact feature descriptors is an effective solution to reduce the transmission data size. Initial research on the topic [7], [19]–[25] demonstrated that the transmission data can be reduced by at least an order of magnitude via extracting compact visual features.

In order to find compact feature descriptions thus reducing transmission bits, various of feature descriptors have been proposed to achieve robust visual content identification under rate constraints. The early-stage keypoint description algorithms assign to each detected keypoint a compact signature consisting of a set of real-valued elements. In [26], an image retrieval system is proposed, based on Harris corner detector

Manuscript received December 15, 2017; revised July 27, 2018; accepted November 3, 2018. Date of publication November 13, 2018; date of current version October 29, 2019. This work was supported in part by NSF of Phase I I/UCRC Center for Big Learning at UMKC under Contract 1747751 and in part by the DDDAS Program of AFOSR. This paper was recommended by Associate Editor Y. Wang. *(Corresponding author: Zhu Li.)*

Z. Zhang, L. Li, and Z. Li are with the Department of Computer Science and Electrical Engineering, University of Missouri-Kansas City, Kansas City, MO 64110 USA (e-mail: zzktb@mail.umkc.edu; lil1@umkc.edu; lizhu@umkc.edu).

H. Li is with the Chinese Academy of Sciences Key Laboratory of Technology in Geo-Spatial Information Processing and Application System, University of Science and Technology of China, Hefei 230027, China (e-mail: lihq@ustc.edu.cn).

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

Digital Object Identifier 10.1109/TCSVT.2018.2881177

1051-8215 © 2018 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

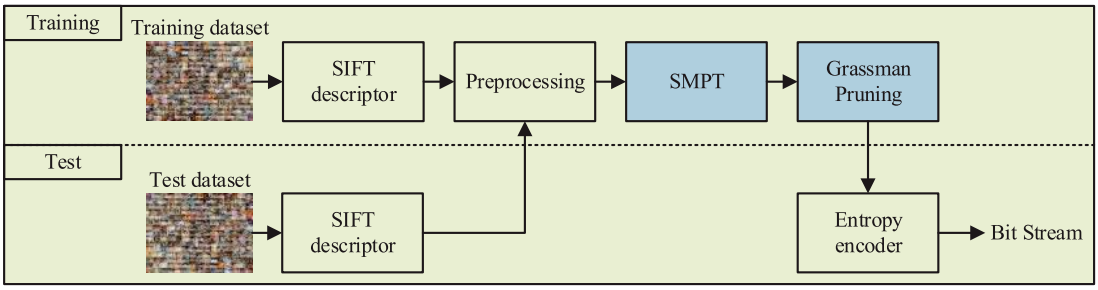


Fig. 1. Architecture of the proposed method. SIFT Manifold Partition Tree (SMPT) is constructed to divide the training SIFT descriptors into small groups wherein each, a transform will be learned with all the SIFT descriptors in that group resulting in a set of local transforms in multiple scales. Grassmann metric is introduced to measure the similarity of two transforms and remove redundant ones. Those remaining optimal transforms are utilized for compression. Conventional entropy encoder generates the bit stream.

and local grayvalue invariants. Such an approach is invariant with respect to image rotation. The work in [27] proposes Shape Context, a feature extraction algorithm that captures the local shape of a patch. Edge detection is firstly performed over a patch surrounding the point (x, y) followed by the radial grid, finally, a histogram centered in (x, y) counts the number of edge points falling in a given spatial bin.

David Lowe introduces Scale Invariant Feature Transform (SIFT) [28] which is the first to achieve scale invariance. SIFT computes for each keypoint a real-valued descriptor, based on the content of the surrounding patch in terms of local intensity gradients. The final SIFT descriptor consists of 128 elements. Given its remarkable performance, SIFT has been often used as starting point for the creation of other descriptors. Inspired by SIFT, Mikolajczyk and Schmid propose Gradient Location and Orientation Histogram (GLOH) [29]. In the context of pedestrian detection, Dalal and Triggs propose Histogram of Oriented Gradients (HOG) [30], a descriptor based on spatial pooling of local gradients. SURF [31] includes a fast gradient-based descriptor. Fan *et al.* propose MROGH [32], a 192-dimensional local descriptor. Along the same line, Girod and co-workers propose rotation invariant features based on the Radial Gradient Transform [33].

To further reduce transmission bits over the wireless network, binary descriptors are proposed. Calonder *et al.* introduce Binary Robust Independent Elementary Features (BRIEF) [34], a local binary keypoint description algorithm. Leutenegger *et al.* propose Binary Robust Invariant Scalable Keypoint (BRISK) [35], a binary intensity-based descriptor inspired by BRISK. Differently from BRIEF, BRISK is able to produce scale-and rotation-invariant descriptors. Similarly to the case of BRISK, Fast REtinA Keypoints (FREAK) [36] uses a novel sampling pattern of points inspired by the human visual cortex, whereas Oriented and Rotated BRIEF (ORB) [37] adapts the BRIEF descriptor, so that it achieves rotation invariance.

However, these descriptors are not compact enough to transmit between remote server and client directly due to their large size [38], [39]. Table I shows the sizes of the descriptors. Take the SIFT as an example, the uncompressed SIFT descriptor is conventionally stored as 1024 bits per descriptor (128 dimensions and 1 byte per dimension). Even a small number of uncompressed SIFT results in tens of KBs.

TABLE I
OVERVIEW OF THE MOST COMMON LOCAL FEATURE DESCRIPTORS

Descriptor	Year	Default size (bytes)
Schmid and Mohr [26]	1999	32
Shape context [27]	2002	144
SIFT [28]	2004	128
GLOH [29]	2005	512
HoG [30]	2005	124
SURF [31]	2006	256
DAISY [40]	2010	400
MROGH [32]	2010	192
BRIEF [34]	2011	64
BRISK [35]	2011	64
ORB [37]	2011	32
FREAK [36]	2012	64

Hence, local feature compression of these raw features is critical for reducing the feature size.

Inspired by these recent developments, CDVS [41] tries to compress SIFT features as well as provides a standardized bitstream syntax to enable interoperability in the context of image retrieval applications and achieves state-of-the-art performance. The local SIFT compression scheme is proposed in [42]. The main idea is to group SIFT descriptors into two groups according to their relative locations and perform linear projection accordingly. Only a subset of descriptors is empirically selected to achieve different coding bit rates. However several drawbacks need to be addressed. First, it is not effective to perform retrieval task in a large-scale dataset by applying the same transform for all the feature descriptors. Second, with the ear of high definition broadcasting and the improvement of hardware, tons of high-resolution images/videos are generated around us. This requires more efficient algorithms to further compress these content.

In this work, we focus on exploiting the intrinsic characteristics of local subspaces in SIFT [28] feature space. As depicted in Fig. 1, a hierarchical partition tree is proposed to divide the whole SIFT dataset into small groups. In each group, a transform will be learned from all these SIFT descriptors in the current group. Grassmann metric is introduced to measure the similarity between every two transforms. A repetitive process of merging similar transforms is devised to remove redundant transforms. Finally, a set of optimal transforms

will be used to compress query descriptors. This work is an extension of the previous work [43], [44] with the following additional contributions:

- We devise a non-linear data partition tree architecture to divide a large-scale training dataset into local patches. The proposed scheme is dedicated to an exploration of capturing the latent characteristics at multiple scales, different numbers of local spaces are generated in each level. The affinity relationship is defined to enable optimization for effective transformation search.
- To search for the optimal local transformations, Grassmann manifold is adopted to prune the data partition tree. We propose a fast search scheme to find the optimal K local transformations from the $\sum_{i=1}^{h} l_i$ candidates. Instead of traversing all the possibilities on the data partition tree, we incorporate Grassmann metric to measure their distinctiveness thus the most representative candidates will be obtained.
- A novel projection and matching metric is devised which involves the local space each feature belongs to. This actually guarantees the fairness of distance computation.
- Theoretical analysis, as well as empirical evidence, are provided to validate the necessity of the proposed data partition tree from the information theory perspective.
- Extensive experiments are performed to evaluate the proposed method, including pairwise matching and large-scale image retrieval. The experimental results show promising results when compared with state of the art methods.

The rest of the paper is organized as follows. Section II introduces the framework of proposed method. The proposed SIFT Manifold Partition Tree (SMPT) is presented in Section III. Weighted Grassmann pruning of SMPT is detailed in Section IV. Experimental setups and results are discussed in Section V. Section VI concludes the work followed by a summary.

II. The Framework

The framework of the proposed method is introduced in this section. As illustrated in Figure 1, CDVS dataset is firstly divided into training and test dataset and each with half the number of images, including both paired images and non-paired images. Each subset in training or test is constructed so that the number of images belonging to each category is proportional to the total amount of images in each category present in the entire database. This paper addresses two innovations in blue background, e.g., SIFT manifold partition tree (SMPT) and Grassmann pruning.

A. Training

All the SIFT descriptors from the training dataset are used for training. A hierarchical multi-level SIFT manifold partition tree is constructed to divide the training samples into small *groups*, *a.k.a. cluster* or *node*. In each group, a transform is learned with all the SIFT descriptors in the current group. In the following section without special explanation, we will refer to *global space* as the transform space learned with the whole dataset without partition, and *local space* as the transform space learned with samples in small groups after SMPT.

Since the total number of local transforms might be very large, in addition, not all these local transforms are optimal transforms, e.g., in extreme cases, the training samples in each local space might be only one which definitely will not be able to train a satisfactory transform. Therefore, there is no need to encode all the local transforms. Grassmann metric is introduced after SMPT to prune all these available local transform candidates. In essence, Grassmann manifold provides a criterion to measure the similarity of two subspaces. In practice, it is ideal to have a group of orthogonal transform bases such that each basis captures the latent characteristics in a unique direction. While in visual query feature compression task, it is also desirable to have a bunch of transforms that they are distinctive to each other. Therefore, under Grassmann metric, we will remove those transforms which are closer to each other and only preserve those have larger Grassmann distances thus they are more likely to capture latent features in a more efficient manner.

The final optimal local transforms after Grassmann pruning will be utilized for compression. Conventional entropy encoder is utilized for encoding to obtain bitstream.

B. Test

As discussed above, half of the randomly-sampled images in CDVS dataset are used for test. Two main experiments have been devised to validate the proposed method, e.g., pairwise matching and large-scale image retrieval experiments. For the former one, given a query SIFT descriptor from the test dataset, it will be assigned to one of those optimal local transforms obtained from the training process. Corresponding local transform will be applied on the query SIFT. It should be noted that the optimal local transforms might be in different level on the SMPT in order to capture hidden characteristics in different scales. That is where the significance of the proposed SMPT structure. Fisher vector [45], [46] aggregation has been applied to the compressed SIFT descriptors to generate the image-level representation for image retrieval experiments.

III. SIFT Manifold Partition Tree

In a large-scale dataset, it is usually not sufficient to capture the intrinsic characteristics in feature space if only one linear transform is applied. Also, due to the variety of dataset sizes, it is difficult to determine the appropriate size of local subspaces. Therefore, it is necessary to design an effective data partition scheme to divide the whole dataset into different levels of small groups. In this work, SIFT Manifold Partition Tree (SMPT) is proposed to divide the global dataset into small groups at different scales.

The SMPT grows in a top-down manner and the number of groups is hierarchically increasing. As the total number of training samples is fixed, different numbers of groups lead to different numbers of samples in each group. This makes it possible to exploit latent discriminative property

at different scales. To assign each sample into a group, conventional aggregation methods are considered, which could be roughly divided into two categories: soft assignment and hard assignment. In our case, it is preferable to use hard assignment as we need to assign a unique local space for each training sample. k-means is adopted in this work as it is able to preserve the geometric structure of the global dataset as well as its nature of simplicity and effectiveness.

In essence, SMPT is constructed by partitioning the large-scale SIFT dataset into small patches followed by proper design of connection relationship. The core of k-means algorithm is an easily-understood optimization problem: given a set of data points (in some vector space), try to position k other points at locations that minimize the (squared) distance between each point and its closest center. We denote the training dataset containing M SIFTs as $X = \{x_m\}$, $m = 1...M$ which has to be partitioned into k clusters. K-means clustering solves

$$\arg\min_{c} \sum_{i=1}^{k} \sum_{x \in c_i} d(x, \mu_i) = \arg\min_{c} \sum_{i=1}^{k} \sum_{x \in c_i} ||x - \mu_i||_2^2 \quad (1)$$

where c_i is the set of points that belong to cluster i. The standard methods for solving the k-means optimization problem are Lloyd's [47] algorithm (a batch algorithm, also known as Lloyd-Forgy [48]).

Since the algorithm stops at a local minimum, the initial position of the clusters is very important. Some common methods to initialize the centroids:

1) Forgy: set the positions of the k clusters to k observations chosen randomly from the dataset.
2) Random partition: assign a cluster randomly to each observation and compute means.

Due to the huge size of MVS datasets, we use the *Forgy* initialization to accelerate convergence. Before partitioning, PCA is applied on the whole SIFT dataset to reduce the dimensionality. There are two reasons for this dimensional reduction operation. Firstly, the lower dimensional local features help to produce compact final image representation as the current in image retrieval. Secondly, applying PCA could help to remove noise and redundancy; hence, enhancing the discrimination [49].

As illustrated in Fig. 2, in each level, the dimension-reduced SIFT descriptors of the whole dataset are divided into small groups via k-means. $n_j^{(i)}$ indicates the j-th node on the i-th level, where $i = 1, \ldots, h$. It should be noted that all small groups from certain level are directly from the whole dataset, not from its parent level, e.g., combining all SIFT samples from any level would constitute the whole dataset. The association is specified in a bottom-up manner. Euclidean distance is calculated between current child-cluster centroid and its parent-cluster centroid. Current child-cluster is associated with the nearest parent-cluster.

At each node, a transform (PCA in this paper) is trained using all full-dimension (128-d) SIFT descriptors. So far, all the transform candidates are obtained.

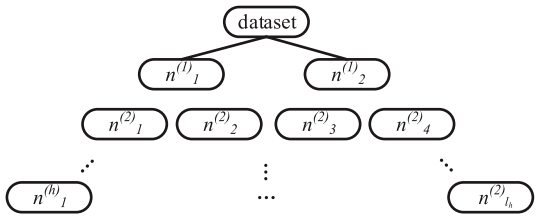


Fig. 2. The SMPT is built in a top-down manner. Each level has l_i clusters obtained from k-means aggregation, where $n_j^{(i)}$ represents the j-th node on the i-th level. The affiliation of child-cluster is processed in a bottom-up manner. Each child-cluster is associated with one parent-cluster according to Euclidean distance.

IV. Weighted Grassmann Pruning of SMPT

As we discussed above, the partition on the global space serves as the first stage of the proposed work. However, leaf nodes may not be the best choice for retrieval because of the following reasons. First, with the increase of the SMPT level, the number of samples associated with each leaf node will decrease. Thus there exist a scenario where the number of samples is not sufficient to train a reliable transform. Take the extreme case as an instance, there will be only one sample in each leaf node when the number of nodes equals the total number of samples in the whole dataset. Second, in practice, it is expensive to encode all available transform candidates. Therefore, it is desirable to devise a scheme to search for a handful of optimal transforms.

We introduce Grassmann manifold [50], [51] into the indexing model for manipulating the leaf nodes derived from the data partition tree. Each point on Grassmann manifold is a subspace by the columns of an orthonormal matrix which is invariant to any basis. The notion of principle angle and Grassmann distance allow us to evaluate the homogeneity of the SIFT feature space. In this section, we first briefly review the Grassmannian metric and related concepts, i.e., principal angles. Then we introduce the details of applying the Grassmann metric to the SMPT.

A. Grassmann Manifold

The Grassmann manifold $G(d, D)$ is the set of d-dimensional linear subspaces of the $\mathbb{R}^D$ [50]. Consider the space $\mathbb{R}_{D,d}^{(0)}$ of all $D \times d$ matrices, i.e., $A \in \mathbb{R}^{D \times d}$. The group of transformation $A = AS$, where S is a $d \times d$ full-rank square matrix, defines an equivalence relation in $\mathbb{R}_{D,d}^{(0)}$.

$$A_1 = A_2 \text{if } \operatorname{span}(A_1) = \operatorname{span}(A_2) \quad \text{where } A_1, A_2 \in \mathbb{R}_{D,d}^{(0)} \quad (2)$$

Therefore, the equivalence classes of $\mathbb{R}_{D,d}^{(0)}$ are one-to-one correspondence with the points on the Grassmann manifold $G(d, D)$, i.e., each point on the manifold represents a subspace. According to the definition, each point on Grassmann manifold is a subspace. Therefore, to measure the distance between two points on the Grassmann manifold is equivalent to measure the similarities between two subspaces. Principle angle [50]–[52] is a geometrical measure between two subspaces. Fig. 3 has shown the relationship between principal

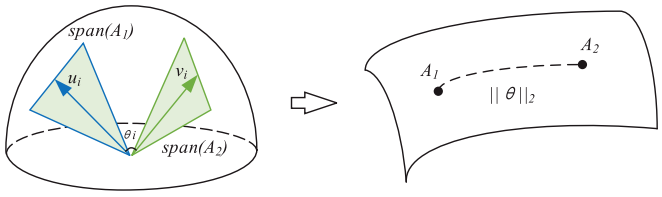


Fig. 3. Principal angles and Grassmann distances. Let $span(A_1)$ and $span(A_2)$ be two subspaces in the Euclidean space $\mathbb{R}^D$ on the left. The distance between two subspaces can be measured by the principle angles $\theta = [\theta_1, \theta_2, \ldots, \theta_m]^T$.

angle and Grassmann manifold. Suppose A_1 and A_2 are two orthonormal matrices $A_1, A_2 \in \mathbb{R}^{D\times d}$ on the Grassmann manifold, the principal angles $0 \leq \theta_1 \leq \cdots \leq \theta_d \leq \pi/2$ between two subspaces $span(A_1)$ and $span(A_2)$, are defined recursively by:

$$\begin{aligned} \cos\theta_k = & \max_{u_k \in span(A_1)} \max_{v_k \in span(A_2)} u_k{}' v_k, \\ & s.t.\ u_k{}' u_k = 1,\ v_k{}' v_k = 1, \\ & u_k{}' u_i = 0,\ v_k{}' v_i = 0,\ (i = 1, ..., k-1) \end{aligned} \quad (3)$$

The vectors $(u_1, u_2, \ldots, u_d)$ and $(v_1, v_2, \ldots, v_d)$ are principal vectors of the two subspaces. θ_k is the kth smallest angle between two principal vectors u_k and v_k.

In literature, there are a variety of methods to compute the principal angles between two subspaces. One numerically stable way is to apply Singular Value Decomposition (SVD) on the product of the two matrices $A_1' A_2$, i.e.,

$$A_1' A_2 = USV' \quad (4)$$

where $U = [u_1, u_2, \ldots, u_d]$, $V = [v_1, v_2, \ldots, v_d]$ and $S = \text{diag}(\cos\theta_1, \ldots, \cos\theta_d)$. The cosine values of principal angles $\cos\theta_1, \ldots, \cos\theta_d$ are known as canonical correlations [52].

B. Subspace Optimization With Grassmann Metric

The distance on Grassmann manifold is defined as follows. A distance is referred to as Grassmann distance if it is invariant under different basis representations. Grassmannian distances between two linear subspaces span(A_1) and span(A_2) can be described by principal angles. The smaller principal angles are, the more similar two subspaces are i.e., the closer they are on the Grassmann manifold.

In literature, various Grassmann distance metrics based on principal angle have been developed for different purposes, e.g., projection, Binet-Cauchy, max correlation, min correlation, Procrustes metric [50]. Since the distance metrics are defined with a particular combination of the principal angles, the best distance depends highly on the probability distribution of the principal angles of the given data. Among all these metrics, max correlation and min correlation only use the maximum and minimum principal angle, respectively, thus may perform less stable when the noise in the data varies. Another criterion for choosing the distance is the degree of structure in the distance. Without any structure, a distance can be used only with a single K-Nearest Neighbor (KNN) algorithm. When a distance having an extra structure such as triangle inequality, for example, we can speed up the nearest neighbor search by estimating lower and upper limits of unknown distances. From this point of view, only Binet-Cauchy metric and projection metric are the most structured metrics as they are induced from a positive definite kernel [50]. In application, they are also the most commonly used Grassmann metrics. Therefore, in the final experimental section, both projection distance and Binet-Cauchy Grassmann distance will be evaluated. The projection Grassmann metric and Binet-Cauchy Grassmann metric can be computed as follows, respectively:

$$d_P(A_1, A_2) = \Big(\sum_{i=1}^{m} \sin^2\theta_i\Big)^{1/2} \quad (5)$$

$$d_{BC}(A_1, A_2) = \Big(1 - \prod_{i} \cos^2\theta_i\Big)^{1/2} \quad (6)$$

We denote the number of nodes in each level as $L = \{l_i\}$, where $i = 1, \ldots, h$. Hence, the total number of available transforms is

$$S = \sum_{i=1}^{h} l_i \quad (7)$$

Grassmann metric is applied to measure the similarities between the candidates. The similarity between every two candidates will be measured. The two transforms with the shortest Grassmann distance indicates they are the most similar, thus should be merged according to the principle of maximizing distinctiveness. Let us denote all the S available transforms as $\{A_1, \ldots, A_S\}$, each of which is trained with all the SIFT descriptors in that node. The number of training SIFT descriptors in each node is $W = \{w_1, \ldots, w_S\}$.

Before the merge, suppose the query SIFT descriptors are assigned to R leaf nodes. A merge cost $\mathcal{L}$ is calculated before each merge operation. Children nodes of which LCA node with lowest merge cost will be merged. First, we need to find out the Lowest Common Ancestor (LCA) of any two existing nodes. As we have R nodes currently, if each two share an LCA node, there would be C_R^2 LCA nodes in total. Let us use G represent C_R^2 for simplicity. The LCA of Node i and Node j can be expressed as follows.

$$\tilde{A}_{ij} = LCA(A_i, A_j),\ i \neq j, \quad \text{and} \quad i, j = 1, \ldots, S \quad (8)$$

It should be noted that not all LCA node has only two children. For example in Figure 4, $\tilde{A}_{8,9} = Node_4$ which only has two children 8 and 9, while $\tilde{A}_{5,8} = Node_2$ has three children 8, 9, and 5. We denote the number of children each LCA node has as $\{t_1, \ldots, t_G\}$. The merge cost of the g-th LCA $\tilde{A}_g$ is calculated as

$$\mathcal{L}_g = \sum_{i=1}^{t_g} w^{(A_i^{(g)})} \times d_{GSM}(\tilde{A}_g, A_i^{(g)}),\ (g = 1, \ldots, G) \quad (9)$$

where $A_i^{(g)}$ is the i-th child of $\tilde{A}_g$, $w^{(A_i^{(g)})}$ is the number of SIFT descriptors in node $A_i^{(g)}$, $d_{GSM}(a, b)$ is the Grassmann distance between transform a and b. All the children of the LCA which has the lowest cost among all the G LCA nodes will be removed and current LCA node will be a new node.

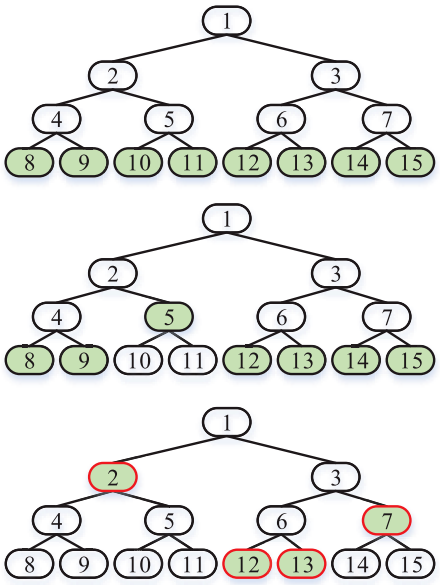


Fig. 4. An example of a 3-*level* SMPT showing the process of pruning. The top figure is the initial state where the query samples are assigned to the leaf nodes. According the merge cost, 10 and 11 are merged to 5; 14 and 15 are merged to 7; 8, 9 and 5 are merged to 2.

Fig. 4 is an example of SMPT showing the procedure of merge similar transforms to achieve the final most distinctive local transforms. At initial status, there are 8 leaf nodes on a 3-*level* SMPT. The objective is to search for 4 most representative transforms. The first step is to find the corresponding LCAs according to Eq. 8. Then calculate all the merge cost according to Eq. 9. Finally, find the LCA with minimum merge cost and all the children corresponding to that LCA will be merged. This process iterates until the number of remaining nodes is 4. By incorporating Grassmann metric, the most distinctive transformations can be achieved.

C. Projection and Matching

Given N query SIFT features $\mathcal{F} = \{f_1, \ldots, f_N\}$ and K optimal local transforms $\{A_1, \ldots, A_K\}$, where $f_n \in \mathbb{R}^{128}$ is a SIFT descriptor. Each query feature in $\mathcal{F}$ is assigned to one of the K nodes. Suppose feature f_n is assigned to transform A_k, the projection is applied as follows.

$$f_n^{(t)} = (f_n - \mu_k) \times A_k \tag{10}$$

where $\mu_k \in \mathbb{R}^{128}$ is the mean of all the training samples in node A_k, $f_n^{(t)}$ is the representation of feature f_n in transform domain.

In transform domain, given a descriptor, its nearest neighbor is searched across the transform domain. The descriptor which has the smallest distance will be marked as its matching pair. Given two projected features $f_{n_1}^{(t)}$ and $f_{n_2}^{(t)}$, different schemes are applied in matching procedure if they are associated with different optimal transforms.

It is not fair if computing the distance between two projected features directly. Because they are in different local spaces, their projections are based on different centers, i.e., centroids are different. Directly computing their distance may exaggerate their real distance. As we do not have the original feature information, so using *pseudo inverse* is an appropriate way to reduce the error introduced in this procedure.

Suppose f_{n_1} is associated with transform A_{k_1} and f_{n_2} is associated with transform A_{k_2}. If they are in the same local space, i.e. $k_1 = k_2$, the distance between these two features is Euclidean distance. If they are in different clusters, i.e. $k_1 \neq k_2$, they have to be converted into a uniform local space via *pseudo inverse*. Empirically, the number of samples in each cluster while training is the factor to determine in which local space to convert to, i.e., the node with more training samples will be selected. Let us use w_{k_1} and w_{k_2} representing the number of training samples used to train A_{k_1} and A_{k_2}, respectively, and $w_{k_1} > w_{k_2}$, the distance is calculated as follows.

$$d_{ij} = \begin{cases} \| f_{n_1}^{(t)} - f_{n_2}^{(t)} \|_{L_2} & \text{if } k_1 = k_2 \\ \| f_{n_1}^{(t)} - f_{n_2}^{(inv)} \|_{L_2} & \text{otherwise} \end{cases} \tag{11}$$

$$f_{n_2}^{(inv)} = (f_{n_2}^{(t)} \times pinv(A_{k_2}) + \mu_{k_2} - \mu_{k_1}) \times A_{k_1} \tag{12}$$

where *pinv* is *pseudo inverse*.

To validate the pseudo-inverse hypothesis, we randomly select 10k matching SIFT pairs from CDVS dataset and check the matching performance with or without pseudo-inverse. The 10k matching SIFT pairs will be assigned to one of 16 leaf nodes of SMPT. Projection with and without pseudo-inverse will be applied to calculate the distance of each SIFT pair in projection domain. Multiple numbers of preserved dimensions are considered. The top-3 accuracy is used to measure the pairwise matching performance. This process is repeated 10 times and different SIFT pairs are used for each time. The accuracy has been listed in Table II.

As can be observed in Table II, the pairwise matching performance of the proposed method with pseudo-inverse is better than that without pseudo-inverse. The superiority in low-dimension cases is more obvious than in high-dimension cases. This might be caused by that in low-dimension circumstances, the distance is more sensitive to noise as more information loss has been induced due to dimensionality reduction.

TABLE II
PAIR-WISE MATCHING RESULTS WITH AND WITHOUT PSEUDO-INVERSE PROJECTION

test #	preserved dimensionality kd							
	w/o pseudo-inverse				w/ pseudo-inverse			
	4	8	16	32	4	8	16	32
1	0.23	0.54	0.69	0.73	0.41	0.71	0.75	0.76
2	0.27	0.57	0.65	0.74	0.39	0.65	0.72	0.80
3	0.19	0.49	0.62	0.72	0.41	0.63	0.73	0.77
4	0.28	0.47	0.59	0.75	0.38	0.69	0.74	0.78
5	0.32	0.51	0.63	0.74	0.42	0.65	0.77	0.75
6	0.24	0.48	0.64	0.76	0.37	0.75	0.73	0.76
7	0.26	0.52	0.66	0.70	0.39	0.68	0.70	0.74
8	0.28	0.51	0.62	0.73	0.39	0.68	0.73	0.76
9	0.30	0.55	0.70	0.75	0.40	0.65	0.75	0.79
10	0.29	0.54	0.67	0.71	0.40	0.74	0.73	0.78
Avg.	0.27	0.52	0.65	0.73	0.40	0.68	0.74	0.77

TABLE III
OVERVIEW OF MPEG CDVS DATASET

Dataset	Category	# images	# matching pairs	# non-matching pairs	# retrieval queries	Mean # relevant images per query
1	Graphics	2500	3000	30000	1500	2
2	Museum Paintings	455	364	3640	364	1
3	Video Frames	500	400	4000	400	1
4	Buildings	14935	4005	48675	3499	4
5	Common Objects	10200	2550	25500	2550	3

V. EXPERIMENTS

Extensive tests have been conducted to evaluate the performance of the proposed method. There are three parts in this chapter: 1) Evaluation framework description. 2) Energy compaction validation and analysis. 3) The final results of pairwise matching and image retrieval experiments along with the analysis of computational complexity.

A. Evaluation Framework

The experiments are performed over CDVS [41] dataset which consists 10,115 matching image pairs and 112,175 non-matching image pairs. The dataset contains images of 5 categories: *graphics, paintings, video frames, buildings* and *common objects*. They were captured with a variety of camera phones and under widely varying lighting conditions. A brief summary is shown in Table III. As stated in Section II, half the number of randomly-sampled images are used for training and the other half for test, including both paired images and non-paired images. The number of images in each category is proportional to the percentage of the number of images in current category to that of the total number of images in the dataset.

Before constructing the SMPT for pairwise matching and image retrieval experiments, we preprocess the data with PCA to reduce the dimensionality. The lower dimensional SIFT features help to produce compact representation. In addition, applying PCA could help to remove noise and redundancy; hence, enhancing the discrimination [49]. A 7-*level* SMPT is constructed with the number of nodes in each level of $L = \{2, 4, 8, 16, 32, 64, 128\}$ e.g., the first level has 2 nodes, the second level has 4 nodes, etc. The connection relationship is processed in a bottom-up manner as illustrated in Fig. 2. In each node on the SMPT, a local PCA transform will be achieved using full-dimensionality (128-d) SIFT descriptors in that node thus resulting in a total number of $\sum_i^7 l_i = 254$ local transforms. The Grassmann pruning is applied to obtain the final $K = \{4, 8, 16\}$ optimal local transforms.

The energy compaction validation experiment provides empirical evidence from the information theory perspective. The objective is to demonstrate the necessity of partitioning the whole dataset into small groups in order to obtain local transforms. The True Positive Rate (TPR) at less than 1% False Positive Rate (FPR) is reported in the pairwise matching experiment. In compliance with CDVS anchor, average bits per descriptor is used to describe the bit stream. SIFT feature extraction and selection is performed in CDVS software, resulting in about 300 SIFT descriptors for each image.

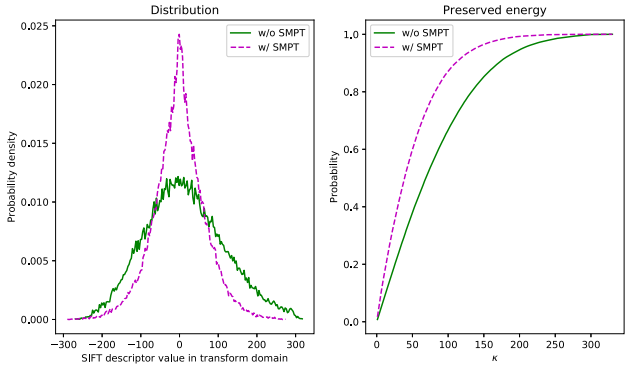


Fig. 5. SIFT descriptor probability distribution and energy preservation plots in transform domain.

Euclidean distance is used to measure the distance in the transform domain. The mean Average Precision (mAP) is used to evaluate the image retrieval performance. The results of each subset are reported for both pairwise matching and image retrieval experiments.

B. Energy Compaction Validation

To study the effects of SMPT, we need to verify the coding efficiency which can be measured by probability distribution histogram. If local transforms are more efficient, the data distribution in transform domain should be more compact, i.e., the probability of a value closer to zero is larger.

We randomly select 60 images from each subset and extract corresponding SIFT descriptors, thus comprises a total of 90k SIFT descriptors. Training samples are achieved by randomly selected by 70%, and the other 30% is used for the test. Before constructing the SMPT, dimensionality reduction technique (e.g., PCA) is applied to reduce the dimension. The dimensionality-reduced SIFTs are used to build a 5-*level* SMPT with the number of nodes in each level $L = \{4, 8, 16, 32, 64\}$. A PCA local transform is trained using all the full-dimensionality (128-d) SIFT descriptors in each node. To remove redundant local transforms, they will be pruned on the Grassmann manifold to achieve final 8 local transforms from $\sum_{i=1}^5 l_i = 124$ available candidates.

After the SMPT model is obtained, the test SIFT descriptors are assigned to one of the 8 optimal nodes according to the Euclidean distance. The test SIFT descriptors in each node will be projected separately using the associated local transform. Fig. 5 shows the probability distribution and preserved energy in the transform domain. It is observed that with SMPT more data value aggregates around zero which is definitely

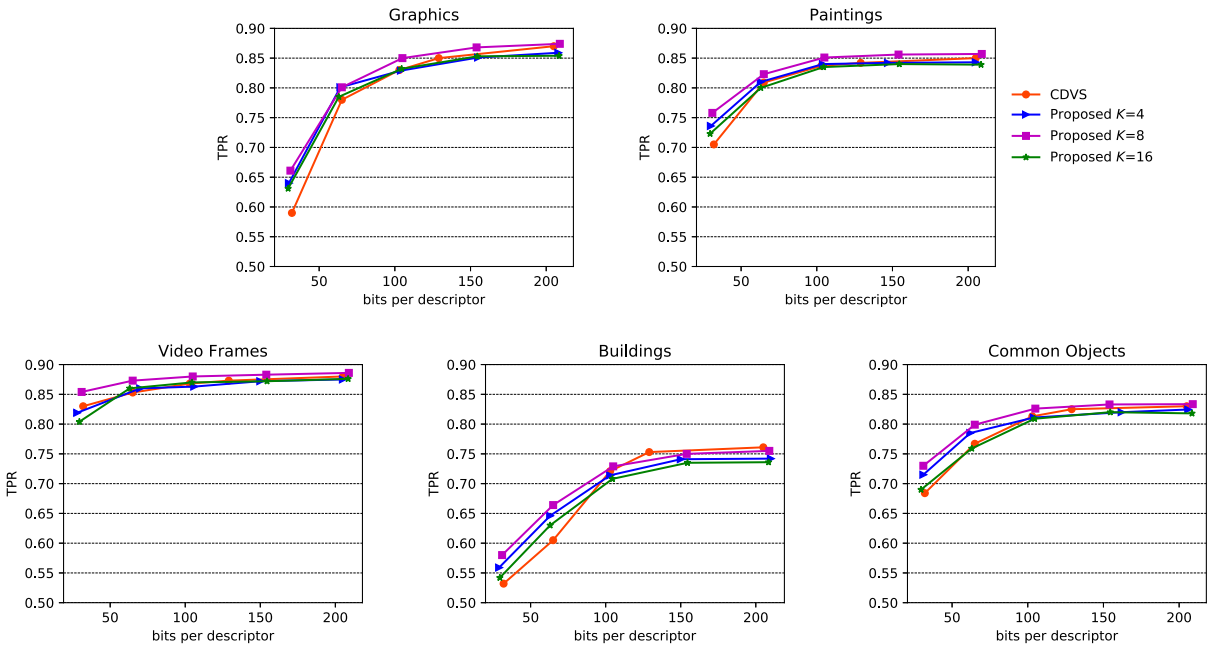


Fig. 6. Results of True Positive Rate (TPR) of the proposed method for SIFT pairwise matching. The proposed method controls bitrate by adjusting the feature dimension in the transform domain, while CDVS provides fixed bitrate configuration. The number of optimal transforms is set as $K = \{4, 8, 16\}$.

TABLE IV
REPEATABILITY COMPARISON OF GRASSMANN PROJECTION DISTANCE AND BINET-CAUCHY DISTANCE WITH CDVS

CDVS		Proposed Projection		Proposed Binet-Cauchy	
bitrate	repeatability	bitrate	repeatability	bitrate	repeatability
32	51.00%	29	52.37%	26	47.48%
-	-	58	68.25%	56	67.92%
65	67.83%	84	72.99%	82	73.72%
-	-	105	76.09%	108	76.68%
103	72.70%	137	78.32%	134	78.20%
-	-	159	79.27%	152	79.54%
129	74.14%	182	80.03%	176	79.54%
-	-	202	80.44%	199	80.76%
205	76.13%	228	80.76%	222	80.92%
-	-	248	80.93%	250	81.17%

beneficial for compression. The preserved energy is calculated by cumulatively summing the probability within range τ away from the origin. When we set the $\tau = 200$, 4.21% more information will be preserved with SMPT than without SMPT.

C. Experimental Results

To compare the difference of Grassmann projection distance and Grassmann Binet-Cauchy distance, repeatability experiments are conducted using SIFT descriptors from all 5 categories. Table IV shows the repeatability results of CDVS, proposed method with Grassmann projection distance and the proposed method with Binet-Cauchy distance, respectively. The bitrate variation of the proposed method is achieved by adjusting the number of the preserved dimensionality of SIFT descriptors. It is observed that the proposed methods perform better than CDVS. Compared with the best repeatability of CDVS, the proposed SMPT with Grassmann projection and Binet-Cauchy achieves about 4.31% and 4.63% improvement with comparable bitrate, respectively. No significant difference has been observed between two Grassmann distance metrics but Binet-Cauchy is slightly better than projection.

Binet-Cauchy is adopted for pairwise matching and image retrieval experiments. Figure 6 and Figure 7 show the pairwise matching and image retrieval results for each subset, respectively. It can be observed that the proposed method achieves the best performance when $K = 8$, where K is the number of optimal transforms after Grassmann pruning. The performance increases when K decreases from 16 to 8, but deteriorates when continue decreasing from 8 to 4. This phenomenon demonstrates that there exists an optimal solution by adjusting the number of transforms. It is a trade-off between the number of training samples to train a transform and the number of transforms utilized to perform projection.

At best performance, it can be seen that the proposed methods perform better than CDVS with only a few exceptions in *buildings* and *Common Objects*. As the proposed method controls bitrate by adjusting the reduced feature dimensionality, it provides more flexibility in practice. In pairwise matching experiments, the proposed method achieves an improvement of 7.1%, 5.3%, 2.4%, 4.8% and 4.6% at lowest bitrate for *Graphics*, *Paintings*, *Video Frames*, *Buildings* and *Common Objects*, respectively. At highest bitrate, an improvement of 0.4%, 0.7%, 0.6% and 0.35% has been observed for *Graphics*, *Paintings*, *Video Frames* and *Common Objects*, respectively. While in *Buildings*, the proposed method is 0.6% worse than CDVS. Similar patterns can been witnessed in image retrieval results. The improvement at lowest bitrate are 4.98%, 3.84%, 2.03%, 3.87% and 4.54% for *Graphics*, *Paintings*, *Video Frames*, *Buildings* and *Common Objects*, respectively. At highest bitrate, the proposed method achieves 1.23%, 0.66%, 0.65% and 0.40% improvement for *Graphics*, *Video Frames*, *Buildings* and *Common Objects*,

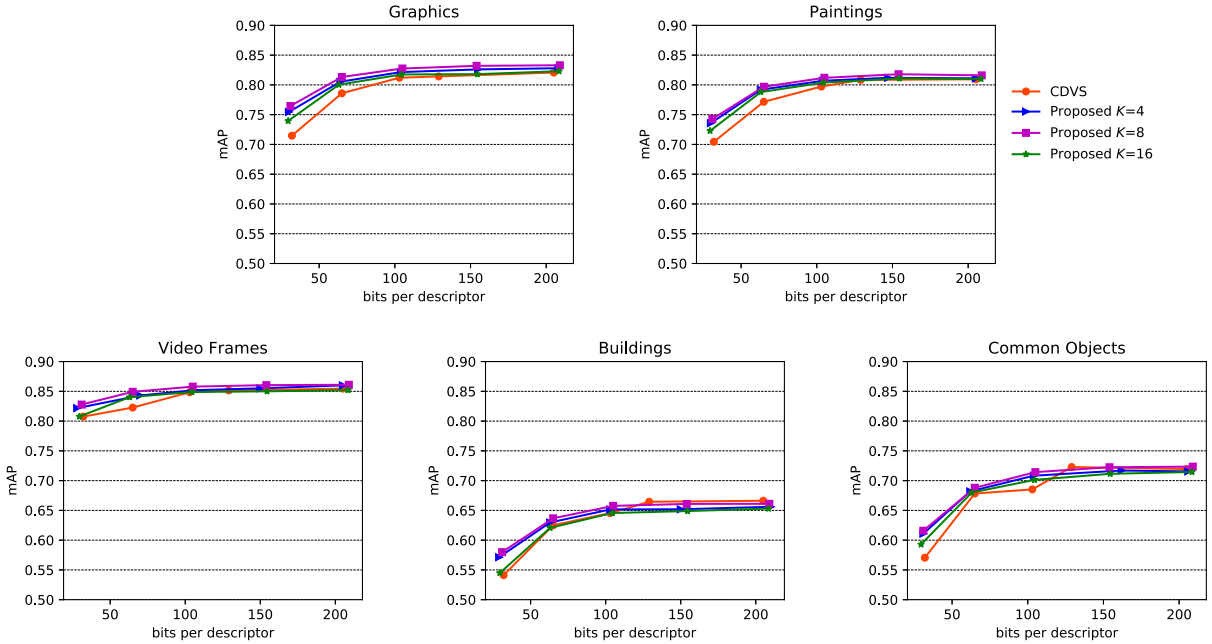


Fig. 7. Results of mean Average Precision (mAP) of proposed method for image retrieval. The proposed method controls the bitrate by adjusting the preserved dimension and CDVS provides fixed descriptor lengths. The number of optimal transforms is set as $K = \{4, 8, 16\}$.

respectively. A tiny drop of 0.51% exists in *Buildings*. *Buildings* and *Common Objects* contain more images and the content vary more substantially in illumination and deformation and contain more undistinguishable distractors than the other three categories. That might be the reason causing relatively lower performance in these two categories.

The experiments are conducted on a Windows PC with Intel Core CPU i7-7700HQ 2.80GHz. The proposed method requires an average of 1.37 milliseconds per query in pairwise matching experiments for highest performance. In image retrieval experiments, an average of 2.49 seconds is required per query at highest performance. The CDVS has been tested using the same dataset and achieves an average of 0.43 milliseconds and 1.84 seconds per query for pairwise matching and image retrieval experiments, respectively. Future work will focus on reducing the computational complexity.

VI. Conclusion

In this paper, we present an effective framework to produce more discriminative image representations for image retrieval using SIFT feature. In the proposed framework, we propose to enhance the discriminative properties in two main steps: i) Use different transforms for query features to capture latent characteristics in multiple scales by constructing SMPT. ii) Prune available local transforms on SMPT to achieve optimal ones by introducing Grassmann metric. Extensive experimental results show that the proposed method achieves promising performance comparing with state of the art. Deep learning-based methods are future research directions as well as a further push to reduce the computational complexity for the proposed scheme.

References

[1] (2017). *Instagram*. [Online]. Available: http://www.internetlivestats.com/one-second/#instagram-band

[2] (2017). *Snapchat*. [Online]. Available: https://www.bloomberg.com/technology

[3] W. Tan, B. Yan, and C. Lin, "Beyond visual retargeting: A feature retargeting approach for visual recognition and its applications," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 11, pp. 3154–3162, Nov. 2017.

[4] G. Takacs *et al.*, "Outdoors augmented reality on mobile phone using loxel-based visual feature organization," in *Proc. 1st ACM Int. Conf. Multimedia Inf. Retr. (MIR)*, New York, NY, USA, 2008, pp. 427–434, doi: 10.1145/1460096.1460165.

[5] S. S. Tsai, D. Chen, J. P. Singh, and B. Girod, "Rate-efficient, real-time cd cover recognition on a camera-phone," in *Proc. 16th ACM Int. Conf. Multimedia (MM)*, New York, NY, USA, 2008, pp. 1023–1024, doi: 10.1145/1459359.1459561.

[6] A. Sarkar, V. Singh, P. Ghosh, B. S. Manjunath, and A. Singh, "Efficient and robust detection of duplicate videos in a large database," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 20, no. 6, pp. 870–885, Jun. 2010.

[7] B. Girod *et al.*, "Mobile visual search," *IEEE Signal Process. Mag.*, vol. 28, no. 4, pp. 61–76, Jul. 2011.

[8] V. R. Chandrasekhar *et al.*, "The stanford mobile visual search data set," in *Proc. 2nd Annu. ACM Conf. Multimedia Syst. (MMSys)*, New York, NY, USA, 2011, pp. 117–122, doi: 10.1145/1943552.1943568.

[9] X. Song, X. Peng, J. Xu, G. Shi, and F. Wu, "Cloud-based distributed image coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 25, no. 12, pp. 1926–1940, Dec. 2015.

[10] Google. (2017). *Google-Goggles*. [Online]. Available: http://www.google.com/mobile/goggles/

[11] B. Erol, E. Antúnez, and J. J. Hull, "HOTPAPER: Multimedia interaction with paper using mobile phones," in *Proc. 16th ACM Int. Conf. Multimedia (MM)*, New York, NY, USA, 2008, pp. 399–408, doi: 10.1145/1459359.1459413.

[12] Amazon. (2007). *Snaptell*. [Online]. Available: http://www.snaptell.com

[13] Layar. (2010). *Layar*. [Online]. Available: http://www.layar.com

[14] Y. Jing *et al.*, "Visual search at pinterest," in *Proc. 21st ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, New York, NY, USA, 2015, pp. 1889–1898, doi: 10.1145/2783258.2788621.

[15] Q. Liu, J. Fan, H. Song, W. Chen, and K. Zhang, "Visual tracking via nonlocal similarity learning," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 10, pp. 2826–2835, Oct. 2018.

[16] C. Lee, C. E. Rhee, and H.-J. Lee, "Complexity reduction by modified scale-space construction in SIFT generation optimized for a mobile GPU," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 10, pp. 2246–2259, Oct. 2017.

[17] J. Jiang, X. Li, and G. Zhang, "SIFT hardware implementation for real-time image feature extraction," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 24, no. 7, pp. 1209–1220, Jul. 2014.
[18] F.-C. Huang, S.-Y. Huang, J.-W. Ker, and Y.-C. Chen, "High-performance SIFT hardware accelerator for real-time image feature extraction," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 3, pp. 340–351, Mar. 2012.
[19] V. Chandrasekhar *et al.*, "Compressed histogram of gradients: A low-bitrate descriptor," *Int. J. Comput. Vis.*, vol. 96, no. 3, pp. 384–399, Jan. 2012, doi: 10.1007/s11263-011-0453-z.
[20] D. M. Chen and B. Girod, "Memory-efficient image databases for mobile visual search," *IEEE Multimedia*, vol. 21, no. 1, pp. 14–23, Jan./Mar. 2014.
[21] D. M. Chen, S. S. Tsai, V. Chandrasekhar, G. Takacs, J. Singh, and B. Girod, "Tree histogram coding for mobile image matching," in *Proc. Data Compress. Conf.*, Mar. 2009, pp. 143–152.
[22] R. Ji *et al.*, "Towards low bit rate mobile visual search with multiple-channel coding," in *Proc. 19th ACM Int. Conf. Multimedia (MM)*, New York, NY, USA, 2011, pp. 573–582, doi: 10.1145/2072298.2072372.
[23] R. Ji *et al.*, "Location discriminative vocabulary coding for mobile landmark search," *Int. J. Comput. Vis.*, vol. 96, no. 3, pp. 290–314, Feb. 2012, doi: 10.1007/s11263-011-0472-9.
[24] J. Lin, L.-Y. Duan, Y. Huang, S. Luo, T. Huang, and W. Gao, "Rate-adaptive compact Fisher codes for mobile visual search," *IEEE Signal Process. Lett.*, vol. 21, no. 2, pp. 195–198, Feb. 2014.
[25] J. Chao, R. Huitl, E. Steinbach, and D. Schroeder, "A novel rate control framework for SIFT/SURF feature preservation in H.264/AVC video compression," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 25, no. 6, pp. 958–972, Jun. 2015.
[26] C. Schmid and R. Mohr, "Local grayvalue invariants for image retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 5, pp. 530–535, 1997.
[27] 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.
[28] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004, doi: 10.1023/B:VISI.0000029664.99615.94.
[29] K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 10, pp. 1615–1630, Oct. 2005, doi: 10.1109/TPAMI.2005.188.
[30] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 1, Jun. 2005, pp. 886–893.
[31] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-up robust features (SURF)," *Comput. Vis. Image Understand.*, vol. 110, no. 3, pp. 346–359, 2008. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1077314207001555
[32] B. Fan, F. Wu, and Z. Hu, "Aggregating gradient distributions into intensity orders: A novel local image descriptor," in *Proc. CVPR*, Jun. 2011, pp. 2377–2384.
[33] G. Takacs, V. Chandrasekhar, S. S. Tsai, D. Chen, R. Grzeszczuk, and B. Girod, "Fast computation of rotation-invariant image features by an approximate radial gradient transform," *IEEE Trans. Image Process.*, vol. 22, no. 8, pp. 2970–2982, Aug. 2013.
[34] M. Calonder, V. Lepetit, C. Strecha, and P. Fua, "BRIEF: Binary robust independent elementary features," in *Computer Vision—ECCV*. Berlin, Germany: Springer, 2010, pp. 778–792, doi: 10.1007/978-3-642-15561-1_56.
[35] S. Leutenegger, M. Chli, and R. Siegwart, "BRISK: Binary robust invariant scalable keypoints," in *Proc. IEEE Int. Conf. Comput.*, Nov. 2011, pp. 2548–2555.
[36] A. Alahi, R. Ortiz, and P. Vandergheynst, "FREAK: Fast retina keypoint," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 510–517.
[37] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "ORB: An efficient alternative to SIFT or SURF," in *Proc. Int. Conf. Comput. Vis. (ICCV)*, Washington, DC, USA: IEEE Computer Society, 2011, pp. 2564–2571, doi: 10.1109/ICCV.2011.6126544.
[38] A. Araujo and B. Girod, "Large-scale video retrieval using image queries," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 6, pp. 1406–1420, Jun. 2018.
[39] L. Sun and G. Liu, "Visual object tracking based on combination of local description and global representation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 21, no. 4, pp. 408–420, Apr. 2011.
[40] E. Tola, V. Lepetit, and P. Fua, "DAISY: An efficient dense descriptor applied to wide-baseline stereo," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 5, pp. 815–830, May 2010.
[41] L.-Y. Duan *et al.*, "Overview of the MPEG-CDVS standard," *IEEE Trans. Image Process.*, vol. 25, no. 1, pp. 179–194, Jan. 2016.
[42] S. Paschalakis *et al.*, *Local Descriptor Compression Proposal*, Standard ISO/IEC JTC1/SC29/WG11 MPEG2012/M25929, Jul. 2012.
[43] Z. Zhang, L. Li, Z. Li, and H. Li, "Visual query compression with locality preserving projection on Grassmann manifold," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2017, pp. 3026–3030.
[44] Z. Zhang, L. Li, and Z. Li, "Visual query compression with embedded transforms on Grassmann manifold," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2017, pp. 393–398.
[45] F. Perronnin and C. Dance, "Fisher kernels on visual vocabularies for image categorization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2007, pp. 1–8.
[46] F. Perronnin, J. Sánchez, and T. Mensink, "Improving the Fisher kernel for large-scale image classification," in *Proc. 11th Eur. Conf. Comput. Vis. IV (ECCV)*. Berlin, Germany: Springer-Verlag, 2010, pp. 143–156. [Online]. Available: http://dl.acm.org/citation.cfm?id=1888089.1888101
[47] S. Lloyd, "Least squares quantization in PCM," *IEEE Trans. Inf. Theory*, vol. 28, no. 2, pp. 129–137, Mar. 1982.
[48] E. W. Forgy, "Cluster analysis of multivariate data: Efficiency versus interpretability of classification," *Biometrics*, vol. 21, no. 3, pp. 768–769, 1965.
[49] T. Hoang, T.-T. Do, D.-K. Le Tan, and N.-M. Cheung, "Selective deep convolutional features for image retrieval," in *Proc. ACM Multimedia*, 2017, pp. 1600–1608.
[50] J. Hamm and D. D. Lee, "Grassmann discriminant analysis: A unifying view on subspace-based learning," in *Proc. ACM 25th Int. Conf. Mach. Learn. (ICML)*, New York, NY, USA, 2008, pp. 376–383, doi: 10.1145/1390156.1390204.
[51] J. Hamm, "Subspace-based learning with Grassmann kernels," Ph.D. dissertation, Dept. Elect. Syst. Eng., Univ. Pennsylvania, Philadelphia, PA, USA, 2008.
[52] T. Wang and P. Shi, "Kernel Grassmannian distances and discriminant analysis for face recognition from image sets," *Pattern Recognit. Lett.*, vol. 30, no. 13, pp. 1161–1165, 2009. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167865509001391

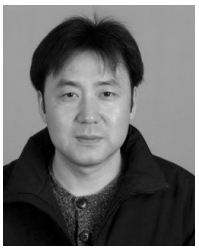
Zhaobin Zhang received the B.S. and M.S. degrees in mechanical electronic engineering from the Huazhong University of Science and Technology, Wuhan, China, in 2012 and 2015, respectively. He is currently pursuing the Ph.D. degree with the Department of Computer Science and Electrical Engineering, University of Missouri-Kansas City. His research interests include machine learning, image/video processing and compression.

Li Li (M'17) received the B.S. and Ph.D. degrees in electronic engineering from the University of Science and Technology of China, Hefei, China, in 2011 and 2016, respectively. He is currently a Post-Doctoral Researcher with the University of Missouri-Kansas City.

His research interests include image/video coding and processing. He was a recipient of the Best 10% Paper Award at the 2016 IEEE Visual Communications and Image Processing Conference.

Zhu Li (M'01–SM'07) is an Associate Professor with the Department of Computer Science and Electrical Engineering, University of Missouri, (UMKC), Kansas City, MO, USA, and the Director of the NSF I/UCRC Center for Big Learning, UMKC. He received the Ph.D. degree in electrical and computer engineering from Northwestern University, Evanston, IL, USA, in 2004. He was a Summer Visiting Faculty at the UAV Research Center, US Air Force Academy, in 2016, 2017, and 2018, respectively. He was a Senior Staff Researcher/Senior Manager with the Samsung Research America's Multimedia Standards Research Lab, Richardson, TX, USA, from 2012 to 2015, the Senior Staff Researcher/Media Analytics Lead with the FutureWei (Huawei) Technology's Media Lab, Bridgewater, NJ, USA, from 2010 to 2012, an Assistant Professor with the Department of Computing, The Hong Kong Polytechnic University, from 2008 to 2010, and a Principal Staff Research Engineer at the Multimedia Research Lab, Motorola Labs, from 2000 to 2008.

His research interests include point cloud and light field compression, graph signal processing and deep learning in the next-gen visual compression, and image/video analysis and understanding. He has 46 issued or pending patents, over 100 publications in book chapters, journals, and conferences in these areas. He serves as a Steering Committee Member for the IEEE ICME since 2015. He is an Elected Member of the IEEE Multimedia Signal Processing, the IEEE Image, Video, and Multidimensional Signal Processing (IVMSP), and the IEEE Visual Signal Processing and Communication Tech Committees. He received the Best Paper Award at the IEEE International Conference on Multimedia and Expo, Toronto, in 2006, and the Best Paper Award (DoCoMo Labs Innovative Paper) from the IEEE International Conference on Image Processing, San Antonio, TX, USA, in 2007. He is the Program Co-Chair of the IEEE International Conference on Multimedia and Expo (ICME) 2019, and Co-Chaired the IEEE Visual Communication and Image Processing 2017. He is an Associate Editor of the IEEE TRANSACTIONS ON MULTIMEDIA, the IEEE TRANSACTIONS ON CIRCUITS & SYSTEM FOR VIDEO TECHNOLOGY, and *Journal of Signal Processing Systems* (Springer) since 2015.



Houqiang Li (M'10–SM'12) received the B.S., M.Eng., and Ph.D. degrees in electronic engineering from the University of Science and Technology of China, Hefei, China, in 1992, 1997, and 2000, respectively. He is currently a Professor with the Department of Electronic Engineering and Information Science, University of Science and Technology of China.

He has authored and co-authored over 100 papers in journals and conferences. His research interests include video coding and communication, multimedia search, and image/video analysis. He was a senior author of the Best Student Paper of the 5th International Mobile Multimedia Communications Conference (MobiMedia) in 2009, and was a recipient of the Best Paper Award for the International Conference on Mobile and Ubiquitous Multimedia from ACM in 2011, the Best Paper Award for Visual Communications and Image Processing in 2012, and the Best Paper Award for International Conference on Internet Multimedia Computing and Service in 2012. He has served on the Editorial Board of the *Journal of Multimedia* since 2009. He served as an Associate Editor for the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY from 2010 to 2013.