VISUAL QUERY COMPRESSION WITH LOCALITY PRESERVING PROJECTION ON GRASSMANN MANIFOLD

Zhaobin Zhang[*] *Li Li*[*] *Zhu Li*[*] *Houqiang Li*[†]

[*] University of Missouri - Kansas City
[†] University of Science and Technology of China

ABSTRACT

For a variety of visual search and visual key points based navigation applications, compression of visual key point features like SIFT is an important part of the overall system that can directly affect the efficiency and latency. In this work, we examine a new approach in visual key points compression, that utilizes subspaces that optimized for preserving key point feature matching properties than the reconstruction performance, and allows for a set of optimal subspaces on Grassmann manifold that can better adapt to the local manifold geometry. The simulation demonstrates that such scheme has very low overhead in signaling subspaces, and has very much improved performance on the repeatability of the keypoint matching subject to bit rate constraints.

Index Terms— Visual query, LPP, Compression, Visual identification, Grassmann manifold

1. INTRODUCTION

Mobile phones and tablets have become pervasive devices which are well designed for visual search. They are equipped with large screens, high-resolution cameras, powerful CPUs along with wireless network function which promotes the emerging industry of query-by-capture applications on mobile devices. A typical client-server image retrieval architecture is presented in Fig.1. Visual descriptors are extracted and compressed on the mobile device. Matching is performed on the server using the transmitted feature data as the query. One of the most challenging work is how to minimize transmission data in order to reduce network latency. Therefore, visual compression is a key process for robust identification of mobile visual content, especially in a very large repository.

To deal with aforementioned problems, various of feature descriptors have been proposed to achieve robust visual content identification under rate constraints, including SIFT [1], SURF [2], GLOH [3], CHoG [4] and RIFF [5]. To further reduce transmission bitrate as well as speed up retrieval process, compact image descriptor has been developed, such as BRIEF [6], a binary descriptor formed from simple intensity difference tests rather than histogram of gradients, PCA-SIFT [7], in which the dimensionality of SIFT feature is reduced with the use of principal component analysis. Inspired by these recent developments, MPEG has carried out the standardization of Compact Descriptors for Visual Search (CDVS) [8].

However, existing methods either lays emphasis on statistic information or reducing dimensionality using a single transform which is not optimal enough to preserve their identification information. They failed to take local subspace relations into consideration such as PCA which is a popular technique used for dimensionality reduction in computer vision. But PCA suffers from a number of shortcomings, such as its implicit assumption of Gaussian distributions [7]. Also, PCA is unable to reveal nonlinear relationships.

In this work, we focus on preserving the local structure and exploring their intrinsic nonlinear properties by incorporating Locality Preserving Projection (LPP) [9] into feature space. The main contributions of this paper are as follows. First, we try to design better transforms which can preserve more local identification relationships in visual query retrieval than PCA. Second, we hierarchically divide the feature spaces into local subspaces to ensure a better adaptation to local geometry. Under the above two contributions, the proposed algorithm is able to find more accurate transform for each query feature as well as reveal nonlinear properties which consequently will improve retrieval accuracy.

The rest of the paper is organized as follows. In section 2 we present LPP embedded compression. In section 3, we discuss the construction of binary tree and introduce Grassmann manifold. Section 4 gives the evaluation results and conclusions are made in section 5.

2. LPP EMBEDDED COMPRESSION

The dimensionality reduction approaches like Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA) have been widely used in computer vision. However, neither of them take local neighborhood relations [9] into consideration which is far more important in visual search than achieving a minimum reconstruction error. Also, they are unable to reveal nonlinear relationships. However, LPP shares

This work is supported by University of Missouri-Kansas City New Faculty Startup Grant.

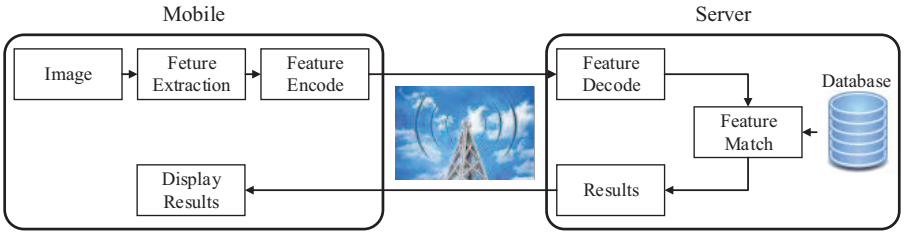

Fig. 1. Mobile server model

many nonlinear projective technique properties and can preserve nearest neighborhood information while rejecting far neighbors which is essentially different from PCA and LDA. Therefore, we involve LPP in our formulation to preserve the nearest neighbors information in the projected space while disregarding far away neighbors [10]. The detailed introduction of LPP will be introduced in the following two subsections.

2.1. Affinity Matrix

A graph incorporating neighborhood information will be constructed. Using the notion of Laplacian of the graph, an affinity matrix will map the data points to a subspace. This linear transformation optimally preserves local neighborhood information in a certain sense [9]. Given a set of query descriptors $W = [x_1, x_2, ..., x_N]$ where $x_i \in \mathbb{R}^n$, the nearest-neighbor relations are represented by their affinity matrix W. The affinity matrix is computed according to the following equation:

$$W_{ij} = \begin{cases} 0 & \text{if } \| x_i - x_j \| > \theta \\ e^{-\|x_i - x_j\|^2/\sigma} & \text{otherwise} \end{cases} \quad (1)$$

σ and θ are the kernel parameter and the cut off threshold respectively. σ controls the shape of the Gaussian function. When σ increases, the affinity matrix entry W_{ij} becomes less sensitive to the distance $\| x_i - x_j \|$. As a rule of thumb, $\sigma = 0.25$ would be a proper choice. As shown in Fig.2, the threshold θ controls the sparsity of affinity matrix. When θ is larger, more neighborhood relations are preserved. However, it is not always beneficial by keeping θ large as far neighborhood relations will also be preserved which may diminish the importance of close connections. When θ is too small, some close neighbor connection information might be lost. Therefore, σ and θ must be properly decided. We optimize θ by finding the value when preserving the most matching pairs in the original feature space. The optimized θ value is listed in Table1.

2.2. LPP Embedding

The essence of LPP compression is to find a basis A so that the original query descriptors $x_i \in \mathbb{R}^n$ is projected into a low dimension feature $y_i = A^T x_i$. This is accomplished by minimizing the following objective function:

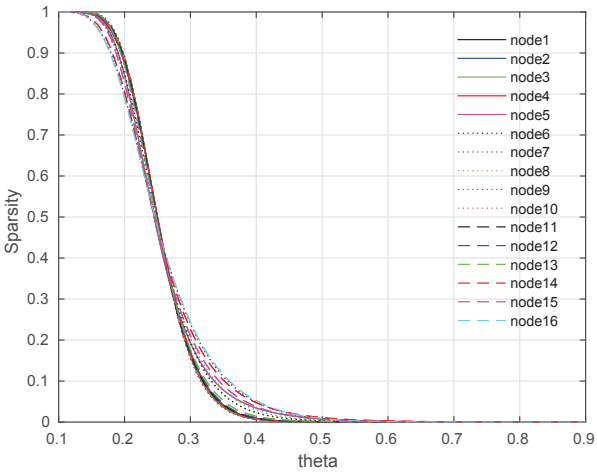

Fig. 2. Sparsity of affinity matrix over multiple θ values

Table 1. Optimized θ value of each node.

Node	1	2	3	4	5	6	7	8
θ	0.17	0.49	0.40	0.10	0.26	0.31	0.39	0.19
Node	9	10	11	12	13	14	15	16
θ	0.23	0.49	0.22	0.66	0.15	0.20	0.51	0.56

$$J(A) = \sum_i^N \sum_j^N \| A^T x_i - A^T x_j \|^2 W_{ij} \quad (2)$$

where A is an $n \times d$ matrix composed of d column vectors $a_1, a_2, ..., a_d$ each of size $n \times 1$ and a_i is orthogonal to a_j, $\forall\, i \neq j$. This objective function will apply heavy penalty if close-by connections are projected far apart. Thus, the resulting A preserves the neighborhood relations in the projected space. The above function can be converted into matrix form [10]. And can be further simplified as:

$$\min_A \{tr[A^T X L X A]\}, \quad s.t.\ tr[A^T X D X A] = 1. \quad (3)$$

Diagonal matrix D is a natural constraint for data points. The bigger $D(k,k)$ is, the more important x_k is. According to [9], we can get A by solving the generalized eigenvalue decomposition problem:

$$XLX^TA = \lambda XDX^TA \tag{4}$$

The column vectors $a_1, a_2, ..., a_d$ which form the basis matrix A correspond to the smallest eigenvalue of the generalized eigenvalue problem.

3. SUBSPACE INDEXING MODEL

To preserve as much local identification information as possible, a single LPP transform is not good enough in capturing all manifold geometry characteristics in feature space. We focus on exploring multiple transforms such that each image descriptor is optimally projected to lower dimensional space meanwhile more identification property is preserved. In this paper, a binary tree is constructed for subspace indexing. When a group of new retrieval features comes, they will be assigned to different local spaces and LPP training will be performed to get the optimal projection for each feature descriptor.

3.1. Binary-Tree Based Indexing

KD-tree is adopted in our work due to its distribution-based property and minimization of quantization error property compared to other local partition methods like Quad-tree. Conventional subspace selection algorithm should be performed before building data partition tree. PCA is an effective approach for indexing.

Given a large-scale dataset which consists of n features, it will be divided into small patches based on a KD-tree. A kd-tree whose height is h has 2^h leaf nodes and in each leaf node, there will be $n/2^h$ features. The covariance information obtained from PCA is utilized in the indexing. We denote the first bases as $A = [a_1, a_2, \cdots, a_d]$. The indexing process is described as follows: 1) project all sample points on the maximum variance basis a_i, find the median value of the projected samples m_1; 2) start from $i = 2$, for each left and right child, project the whole collection of data along the i-th maximum variance basis a_i, find the median value m_i, and split all the children at m_i; 3) increment i and repeat step 2 until some predefined criteria for number of levels, or the number of samples in the leaf nodes is satisfied. However, leaf nodes may not be the most representatives for visual query indexing. First, as the height of the KD-tree increases, the number of samples in each leaf node may be insufficient to get satisfactory projection. On the contrary, if the number of samples in each leaf nodes is too large, it will be more computationally consuming.

3.2. Grassmann Manifold Distance Constraints

In order to find the optimal level of the binary tree such that each descriptor can be optimally projected to low dimensional feature space, we introduce Grassmann distance to control growing of the binary tree.

The Grassmann manifold $G(m, D)$ is the set of m-dimensional linear subspaces of the $\mathbb{R}^D$ [11]. An element on $G(m, D)$ can be represented by an orthonormal matrix A of size D by m such that $A'A = I_m$, where I_m is the $m \times m$ identity matrix. Distance between two Grassmann manifold can be expressed as a polynomial in terms of principal angle. Suppose A_1 and A_2 are two orthonormal matrices of size $D \times m$, the principal angles $0 \leq \theta_1 \leq \cdots \leq \theta_m \leq \pi/2$ between two subspaces $span(A_1)$ and $span(A_2)$, are defined recursively by:

$$\begin{aligned} &\cos\theta_k = \max_{u_k \in span(A_1)} \max_{v_k \in span(A_2)} {u_k}' v_k, \\ &s.t. \quad {u_k}' u_k = 1, \quad {v_k}' v_k = 1, \\ &\qquad {u_k}' u_i = 0, \quad {v_k}' v_i = 0, \quad (i = 1, ..., k-1) \end{aligned} \tag{5}$$

The Binet-Cauchy Grassmann distance is adopted in our method which can be computed as follows:

$$\begin{aligned} d_{BC}(A_1, A_2) &= (1 - \prod_i \cos^2\theta_i)^{1/2} \\ &= (1 - det^2(A_1^T A_2))^{1/2} \end{aligned} \tag{6}$$

Global dataset will be first divided into small patches generating a hierarchy indexing tree. When a new query arrives, the features will be assigned to different nodes with Binet-Cauchy Grassmann distance constraints. For example in Fig.4, the node at which will be further divided is:

$$I_i = \arg\max_i \{d_i\} \tag{7}$$

where

$$\begin{aligned} d_1 &= d_{BC}(A_3, A_1) + d_{BC}(A_4, A_1) \\ d_2 &= d_{BC}(A_5, A_2) + d_{BC}(A_6, A_2) \end{aligned} \tag{8}$$

Training will be performed on each node thus that each feature will have a optimized transform. Fig.3 shows the flowchart of proposed subspace indexing model.

Since we need to select a transform from N transforms for each visual feature, a transform index is needed to be signaled to the decoder. However, there is no need for us to signal a transform index for each visual feature. Different from the natural images, the index of the visual features can be encoded in a very efficient way due its disorder characteristic. We can divide all the visual features into K categories according the transform they used. Then for each category, only $log_2(N)$ bits is needed to signal the transform. Therefore, for all the visual features, only $K \times log_2(N)$ bits are needed to signal the transform index.

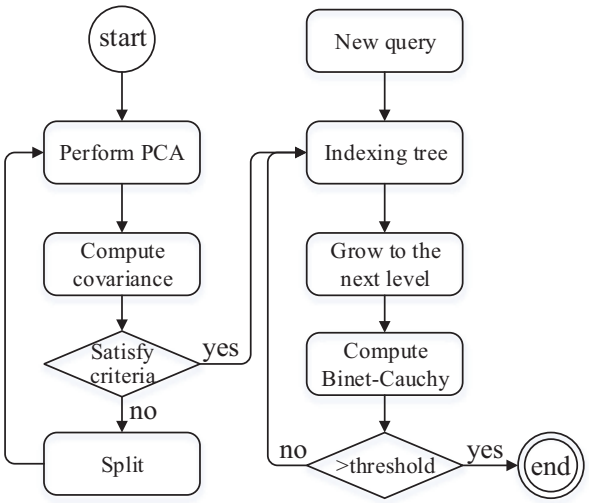

Fig. 3. Subspace indexing model with Grassmann distance constraints

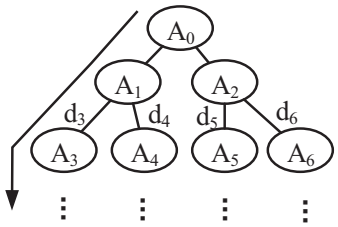

Fig. 4. Binary tree growing scheme with Binet-Cauchy Grassmann distance constraints.

4. EXPERIMENTS

Experiments are conducted on CDVS [8] dataset which consists of a wide variety of items including paintings, book covers, CDs, frames captured from video clips and buildings. A brief introduction of CDVS dataset is described in Table2.

SIFT feature from CDVS dataset is adopted to implement the experiments. SIFT features will be assigned to local subspaces. LPP transforms will be computed for each node. In our implementation, we control the number of final nodes is 16. In each node, we randomly select 4000 samples for training LPP, 4000 samples for evaluation. $d0$ is 0.25 and θ is optimized according to Table1. Multiple dimensions are evaluated and the experiment results are given out in Fig.5. The red lines represent proposed method and the blue lines represent PCA. The results demonstrate a significant improvement in repeatability with the same bitrate constraint. Moreover, we also evaluated on multiple image pairs, the overall average repeatability with the same bitrate constraint has been improved by 23.84%.

Table 2. A brief view of CDVS dataset

Dataset	MP	NMP
1. CDs, DVDs, books, business cards (Mixed text + graphics)	3000	29,903
2. Museum paintings	363	3639
3. Video frames	399	3999
4. Landmarks and buildings	1789	17,949
5. Common object or scenes	2549	21,307

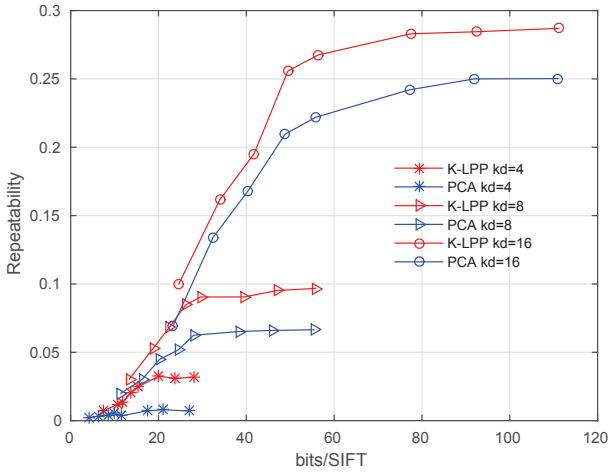

Fig. 5. Average repeatability of K-LPP and PCA with 16 nodes.

5. CONCLUSION

Although lots of approaches have been proposed to put visual query compression technique forward, seldom of them take local geometric information into consideration. Therefore, this paper proposes using multiple projective transforms instead of single transform. LPP is incorporated to preserve more local identification information and try to explore nonlinear matching relationships. We optimize training performance by dividing global training samples into small patches with Grassmann distance constraints. The proposed method is evaluated in MPEG CDVS dataset. The experimental results show that the proposed method which using multiple transforms with optimization on Grassmann manifold can outperform traditional PCA in repeatability under the same bitrate.

6. REFERENCES

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

[2] Herbert Bay, Andreas Ess, Tinne Tuytelaars, and

Luc Van Gool, "Speeded-up robust features (surf)," *Computer Vision and Image Understanding*, vol. 110, no. 3, pp. 346 – 359, 2008, Similarity Matching in Computer Vision and Multimedia.

[3] Krystian Mikolajczyk and Cordelia Schmid, "A performance evaluation of local descriptors," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 10, pp. 1615–1630, Oct. 2005.

[4] Vijay Chandrasekhar, Gabriel Takacs, David M. Chen, Sam S. Tsai, Yuriy Reznik, Radek Grzeszczuk, and Bernd Girod, "Compressed histogram of gradients: A low-bitrate descriptor," *International Journal of Computer Vision*, vol. 96, no. 3, pp. 384–399, 2012.

[5] G. Takacs, V. Chandrasekhar, S. Tsai, D. Chen, R. Grzeszczuk, and B. Girod, "Unified real-time tracking and recognition with rotation-invariant fast features," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, June 2010, pp. 934–941.

[6] Michael Calonder, Vincent Lepetit, Christoph Strecha, and Pascal Fua, *BRIEF: Binary Robust Independent Elementary Features*, pp. 778–792, Springer Berlin Heidelberg, Berlin, Heidelberg, 2010.

[7] Yan Ke and Rahul Sukthankar, "Pca-sift: A more distinctive representation for local image descriptors," in *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Washington, DC, USA, 2004, CVPR'04, pp. 506–513, IEEE Computer Society.

[8] L. Y. Duan, V. Chandrasekhar, J. Chen, J. Lin, Z. Wang, T. Huang, B. Girod, and W. Gao, "Overview of the mpeg-cdvs standard," *IEEE Transactions on Image Processing*, vol. 25, no. 1, pp. 179–194, Jan 2016.

[9] Xiaofei He and Partha Niyogi, "Locality preserving projections," in *In Advances in Neural Information Processing Systems 16*. 2003, MIT Press.

[10] Xin Xin, Zhu Li, and Aggelos K. Katsaggelos, "Laplacian embedding and key points topology verification for large scale mobile visual identification," *Signal Processing: Image Communication*, vol. 28, no. 4, pp. 323 – 333, 2013, Special Issue: VS&AR.

[11] Jihun Hamm and Daniel D. Lee, "Grassmann discriminant analysis: A unifying view on subspace-based learning," in *Proceedings of the 25th International Conference on Machine Learning*, New York, NY, USA, 2008, ICML '08, pp. 376–383, ACM.