SORTING LOCAL DESCRIPTORS FOR LOW BIT RATE MOBILE VISUAL SEARCH

Jie Chen[⋆] *Ling-Yu Duan*[⋆] *Rongrong Ji*[⋆†] *Hongxun Yao*[†] *Wen Gao*[⋆†]

[⋆]Institute of Digital Media, Peking University, Beijing 100871, China
[†]Visual Intelligence Laboratory, Harbin Institute of Technology, Heilongjiang, 150001, China

ABSTRACT

State-of-the-art mobile visual search systems put emphasis on developing compact visual descriptors [4][6], which enables low bit rate wireless transmission instead of delivering an entire query image. In this paper, we address the orderless nature of the transmission set of query descriptors . We propose to adapt the orders of local descriptors in transmission, which subsequently yields more consistent statistic distributions in each feature dimension towards more efficient residual coding based compression. Our scheme further enables lossy sorting by an adaptive quantization strategy within each feature dimension, which largely improves the compression rates of the residual coding in each dimension. We show that the performance degeneration of such lossy sorting is acceptable in our mobile landmark search applications. Our approach's effectiveness and efficiency is demonstrated via extensive experimental comparisons to state-of-the art works in both mobile visual descriptors [2][4] and compact image signatures [5].

Index Terms— mobile visual search, compressive visual descriptors, sorting, lossy compression

1. INTRODUCTION

With the popularization of camera embedded mobile devices and wireless Internet services, nowadays there is an emerging potential in mobile visual search and related applications. Generally speaking, most state-of-the-art mobile visual search systems follow a client-server architecture. The remote server maintains a large-scale photo database, where photos are bound with related information, e.g. locations, touristic and commercial information. To enable the scalability of the ranking system, inverted indexing is usually adopt, typically built upon scalable bag-of-words models, such as works in [1][8][9]. In online search, the mobile user takes a photo as a query, which is transmitted to the remote server to identify its matching photos. Consequently, the best matched photos, together with their bounded information are returned to the mobile user as the search results.

In a typical scenario, the query photo transmission from the mobile devices to the remote server is often over a relatively slow, bandwidth-constrained wireless network. In such case, sending the entire image is often time consuming, which largely degenerates the user experiences in most existing mobile visual search applications. Coming with the ever growing computational power of the mobile devices, recent works have directly extracted visual descriptors on the mobile devices for a low cost transmission. In such case, the compression rates of the traditional compressive local descriptors (e.g. SURF [11] and PCA-SIFT [12]) are still not sufficient for the low bit rate wireless transmission.

Recent works have focused on extreme low bit rate visual descriptors such as [2][3][4][6], which are specially designed for the mobile visual search scenarios. For instance, Chandrasekhar et al. proposed a Compressed Histogram of Gradient (CHoG) [4] for compressive local feature description, which adopts both Huffman Tree and Gagie Tree to compress each local feature into approximate 50 bits. The work in [6] compressed the SIFT descriptor with Karhunen-Loeve Transform, resulted into approximate 2 bits per SIFT dimension. Tsai et al. [10] proposed to transmit the spatial layouts of interest points to improve the discriminability of CHoG. Two recent works also transmit the bag-of-features histogram [2][3] instead of the local feature descriptors: Chen et al. proposed to compress the sparse bag-of-features [2] by encoding position difference of non-zero bins. It produced an approximate 2KB code per image for a vocabulary with 1M words. The work in [3] further compressed the inverted indices of vocabulary tree [1] with arithmetic coding, which further reduces the maintaining memory cost in the mobile device.

In this paper, we focus on extracting and transmitting the local descriptor set through the wireless network. Especially, while such local descriptors are independent with each other, we propose to sort their transmitting orders to facilitate even lower bit rate wireless transmission with residual coding comparing with the state-of-the-art works [4][6]. To the best of our knowledge, this serves as the first work to consider the descriptor orders in the context of mobile visual search.

The rest of this paper is organized as follows: Section 2 introduces our proposed local descriptor sorting scheme, with a subsequent residual coding for low bit rate wireless query transmission. Section 3 presents our mobile landmark search system, with quantitative comparisons to state-of-the-arts [2][4][5]. We conclude this paper in Section 3 and discuss our future research directions.

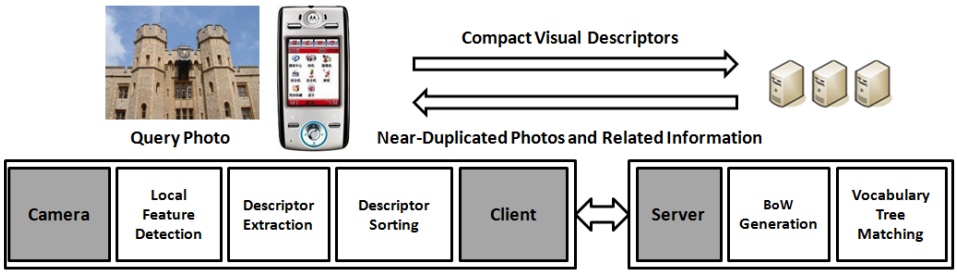

Fig. 1. The proposed low bit rate mobile visual search framework via sorting local feature descriptors [4][7].

2. SORTING LOCAL DESCRIPTORS WITH RESIDUAL CODING

Figure 1 shows the systematic framework of the proposed low bit rate mobile visual search system. Similar to state-of-the-art works in generic and compact visual descriptors [2][3][4][6], our framework emphasizes on directly extracting compact local feature descriptors on a mobile device, then transmitting more reduced-sized query data over the wireless network. Comparing with existing works [2][3][4][6], our contribution lies in sorting local descriptors followed by residual coding to further reduce the transmission rate. Since the local descriptors are delivered in orderless manner through wireless network , their sorting can be performed in an arbitrary order. We argue that sorting these descriptors can bring about potential benefits for subsequent coding methods such as residual coding or wavelet transformation to further encode the entire data amount to transmit.

The Proposed Mobile Visual Search Pipeline: Given a query photo $\mathbf{I}_q$, a mobile device extracts a set of local descriptors as $\mathbf{S}_q = [S_1^q, S_2^q, ..., S_J^q]$. Once this feature set arrives at the remote server, Scalable Vocabulary Tree (SVT) [1] is employed to find the near-duplicated photos (such as landmarks or objects) in an efficient manner. SVT is well exploited in state-of-the-art works [2] [3][8][9], which uses hierarchical k-means to partition local descriptors into quantized codewords. SVT quantizes $\mathbf{S}_q$ by traversing in the vocabulary hierarchy to find the nearest codeword, which converts $\mathbf{S}_q$ into a Bag-of-Words (BoW) signature $\mathbf{V}_q = [V_1^q, V_2^q, ..., V_M^q]$. In each codeword's inverted index, all co-indexed photos are included into the similarity ranking to search near-duplicated photos.

Sorting Local Descriptors From the above pipeline, search is independent of the sending order of these local descriptors. Therefore, we propose to learn an ordering function $[R_1, R_2, ...R_J]$ for $\mathbf{S}_q$ to sort them. Subsequently, we only transmit $\mathbf{S}_q$ via the following residual coding:

$$S_{R_1}, Residual(S_{R_2}), ..., Residual(S_{R_J}) \quad (1)$$

where the $Residual(S_{R_i})$ denotes the residual coding from S_{R_i} to its previous descriptor $S_{R_{i-1}}$, which is calculated as:

$$Residual(S_{R_i}) = [Dis(s_i^1, s_{i-1}^1), ..., Dis(s_i^K, s_{i-1}^K)] \quad (2)$$

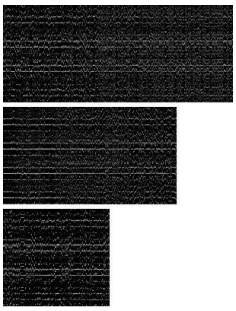

Fig. 2. The visualized example of orderless SIFT sets for three photo. Different query produces different amount of SIFT, hence the horizontal length is not identical. Each row's height is identical, denoting 128 SIFT dimensions. Brighter points denote highes values.

$k \in [1, K]$ denotes the dimension of a given local descriptor, say 128 for SIFT. We select each new S_{R_i} that minimizes the distance to its former descriptor $S_{R_{i-1}}$. Fixing the cost of storing the first local descriptor S_{R_1}, the overall cost with current ordering is given by:

$$Lost_{Sort} = \sum_{j=1}^{J} Residual(S_{R_i}) \quad (3)$$

Greedy and Lossy Learning: Learning the sorting orders by Equation 3 is NP hard. We resort to its greedy version: At the selection of the ith local descriptor, we attempt to find out the descriptor that has the minimal cost of $Residual(S_{R_i})$ using Equation 2. In such case, learning is efficient, with the complexity of only $O(J^2)$ to learn the best order. Let's step forward by look forward the N local descriptors $[S'_1, S'_2, ..., S'_N]$, which means at each time we select a batch of local descriptors as:

$$\mathbf{S}_{Patch} = \arg min_{\mathbf{S}'_N}(Residual(S'_1) + \sum_{n=2}^{N} Residual(S'_n)) \quad (4)$$

Such selection leads to in total C_J^N magnification of computational cost to accomplish its greedy selection.

We further consider the case of lossy learning, which means that the coding residual for a given dimension is fur-

ther quantized into M fractions. As a result, we have the following simplification:

$$Residual(S_{R_i}) = [Q_k(Dis(s_i^k, s_{i-1}^k))]_{k=1}^{K} \quad (5)$$

3. MOBILE SEARCH SYSTEM IMPLEMENTATIONS AND QUANTITATIVE RESULTS

Data Collection: We collected over 10 million geo-tagged photos from photo sharing websites of Flickr and Panoramio. Our dataset covers typical areas including Beijing, New York City, Lhasa, Singapore, and Florence.

Ground Truth Labeling: First, within each city, we selected the top 30 most densely geographical regions, as well as 30 random geographical regions. Second, as manually identifying all related photos from an identical landmark is intensive, for each of these 60 regions, we ask our volunteers to manually identify one or more dominant landmark views. All their near-duplicated landmark photos are labeled in its current region and all the geographical nearby regions. Finally, we randomly sample 5 images from each region to form the ground truth. Each query and their returning results are used to test the retrieval performance of different methodologies. So that we generate 300 user query logs and their ground truth for each city (in total 300×5 queries)

Parameters Tuning and Evaluation: For the landmark photo collection within each city, we extract both SIFT [7] and CHoG [4] features from each photo. We build a Scalable Vocabulary Tree model [1] to produce the initial Vocabulary $\mathbf{V}$, which generates a bag-of-words signature $\mathbf{V}_i$ for each database photo $\mathbf{I}_i$. We denote the hierarchical level as H and the branching factor as B, which gives at most $M = B^H$ words at the finest level. We stop the quantization division once there are less than 1,000 SIFT features in a word (node). Typically, we have $H = 6$ and $B = 10$, producing approximate 100,000 codewords.

We use Mean Average Precision at N ($MAP@N$) against the compression rate to evaluate our system performance. MAP@N represents the mean precision of a batch of queries, each of which reveals its position-sensitive ranking precision within the top N ranking positions.

Baselines: (1) *Transmitting the original query photo*: As the most common strategy in most existing mobile visual search scenario, we resort to JPEG compression to encode an entire query photo. It serves as the first baseline with the highest coding length in our comparisons. (2) *Transmitting SIFT Descriptor Sets*: Transmitting the original local descriptor set serves as a better solution, comparing with transmitting the entire query photo. Since SIFT [7] feature is widely adopted, we use the SIFT descriptor set as our another baseline. (3) *Transmitting CHoG Descriptor Sets*: As an alternative approach, we replace the SIFT descriptor with CHoG [4], which produces a more compact local descriptor with approximate 50 dimensions. Although it does not bring about theoretical improvements in our subsequent compression rate, it can well reduce the storage/memory cost in a mobile device. Also, incorporating CHoG instead of SIFT is shown to yield better ranking MAP. Hence, we choose CHoG as our final local feature descriptor in our system implementation.

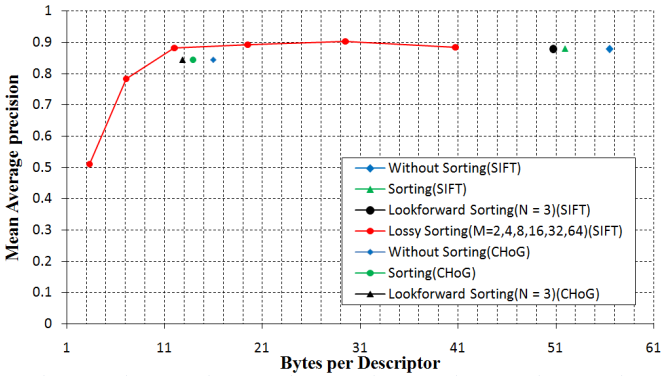

Fig. 3. Compression rate and ranking distortion comparing with [2][4][5] using our ground truth query set. The six points (left to right) in Lossy Sorting correspond to $M = 2, 4, 8, 16, 32, 64$.

Rate Distortion Analysis: We give the rate distortion analysis of our approach in comparisons to alternative methods and state-of-the-art works in [2][4][5]. It is obvious that our schemes have achieved the best performances, as shown in Figure 3. It is obvious that our scheme has not serious lost of MAP before quantization of $M = 8$, which means the SIFT descriptor used here is indeed redundant and compressible.

Lossy Sorting Performance Degeneration: We further investigate the performance degeneration of using our lossy sorting strategy. In Figure 4, we investigates the MAP degeneration using different quantization scheme, where $Q_i = M$ denotes we subdivide each local descriptor dimension into M fractions, and the setting of $M = [2, 4, 8, 16, 32, 64]$ denotes the different division scheme in lossy quantization.

Generally speaking, the curves ($M = 16$, 32, and 64) in Figure 4 is smooth with less MAP degeneration (vertical viewpoint), until $M = 8$. The figure shows the compression rates with different quantization schemes for $M = [2, 4, 8, 16, 32, 64]$ to compare with the lossless compression scheme. Also, since before $M = 8$ the distortion is not obvious, we have selected $M = 16$ in our subsequent case study.

Case Study: Figure 5 shows several groups of exemplar transmission rates, in which we show the given landmark query as well as its corresponding transmission rate. While in general the coding length is variable (corresponding to different landmark queries), it is obvious that our approach can give the most compact visual description length comparing with directly sending the unordered descriptor sets. Finally, the sorted local descriptor set corresponding to Figure 2 is given in Figure 6

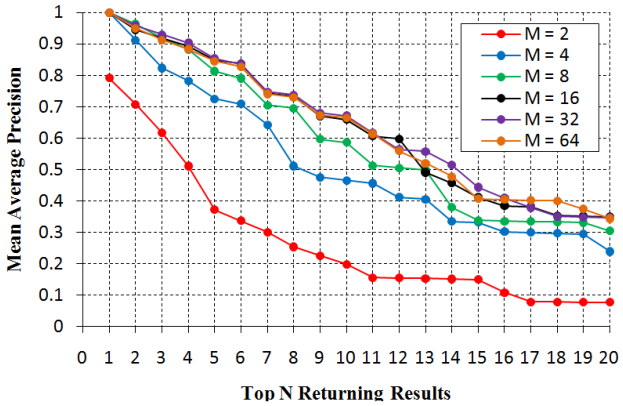

Fig. 4. The retrieval MAP lost of different lossy sorting schemes by quantization scales of $M = [2, 4, 8, 16, 32, 64]$.

Fig. 5. Case study of transmission rates (Bytes Per Local Descriptor) for representative landmark queries in Beijing. Each photo on the left is the query, its right histogram is the transmission rate of transmitting the unordered local descriptors (Green), lossless sorting (Blue), and lossy sorting $M = 16$ (Red), respectively.

4. CONCLUSIONS AND FUTURE WORKS

We have studied the feasibility of sorting the local descriptors at a mobile device, which enables low bit rate wireless transmission in the context of mobile visual search. Such sorting is able to facilitate the subsequent residual coding for a compressive transmission of the original orderless local descriptor set. We further enable lossy sorting by adaptively quantizing the residual values into different fractions, which achieves even higher compression rate without serious lost in retrieval performance. We reported the superior performance over the state-of-the-art work in compact descriptors for visual search. [4]. Currently, the residual coding only serves as an exemplar and suggestive solution to compress our "smooth" (after sorting) local descriptor sets. Our future work includes more investigation on different coding schemes, such as wavelet or DCT transforms, to further compress the proposed local descriptor sorting scheme.

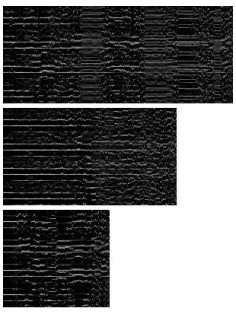

Fig. 6. The 3 corresponding local descriptor sets of Figure 2 after the lossless sorting (lookforward step $N = 1$).

5. ACKNOWLEDGEMENTS

This work was supported in part by grants from the Chinese National Natural Science Foundation under contract No. 60902057, in part by the National Basic Research Program of China under contract No. 2009CB320902, and in part by the CADAL Project Program.

6. REFERENCES

[1] Nister D. and Stewenius H. Scalable recognition with a vocabulary tree. *CVPR*. 2006. 1, 2, 3

[2] Chen D., Tsai S., and Chandrasekhar V. Tree histogram coding for mobile image matching. *DCC*. 2009. 1, 2, 3

[3] Chen D., Tsai S., Chandrasekhar V., Takacs G., Vedantham R., Grzeszczuk R., and Girod B. Inverted index compression for scalable image matching. *DCC*. 2010. 1, 2

[4] Chandrasekhar V., Takacs G., Chen D., Tsai S., Grzeszczuk R., and Girod B. CHoG: Compressed histogram of gradients a low bit-rate feature descriptor. *CVPR*. 2009. 1, 2, 3, 4

[5] Jegou H., Douze M., Schmid C., Perez P. Aggregating local descriptors into a compact image representation. *CVPR*. 2010. 1, 3

[6] Chandrasekhar V., Takacs G., Chen D., Tsai S., Singh J., and Girod B. Transform coding of image feature descriptors. *VCIP*. 2009. 1, 2

[7] Lowe D. G. Distinctive image features from scale-invariant keypoints. *IJCV*. 2004. 2, 3

[8] Irschara A., Zach C., Frahm J., Bischof H. From structure-from-motion point clouds to fast location recognition. *CVPR*. 2009. 1, 2

[9] Schindler G. and Brown M. City-scale location recognition. *CVPR*. 2007. 1, 2

[10] Tsai S., Chen D., Takacs G., Chandrasekhar V. Location coding for mobile image retrieval. *MobileMedia*. 2010. 1

[11] Bay H., Tuytelaars T., and Van Gool L. SURF: Speeded up robust features. *ECCV*. 2006. 1

[12] Ke Y. and Sukthankar R. PCA-SIFT: A more distinctive representation for local image descriptors. *CVPR*. 2004. 1