A LOW BIT RATE VOCABULARY CODING SCHEME FOR MOBILE LANDMARK SEARCH

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

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

ABSTRACT

We present a low bit rate vocabulary coding scheme in the context of mobile landmark search. Our scheme exploits location cues to boost a compact subset of visual vocabulary, which is discriminative for visual search and incurs low bit rate query for efficient upstream wireless transmission. To validate the coding scheme, we have developed mobile landmark search prototype systems within typical areas including Beijing, New York City, Lhasa, Singapore, and Florence. Our system maintains a single vocabulary in a mobile device, which can be efficiently adapted with the location information of city-scale mobile users. Thus multiple downloading of large vocabulary is completely avoided for normal city tourists. In landmark search domain, we have reported superior performance over the state-of-the art works in compact image descriptors or signatures [2][4][5].

Index Terms— landmark search, mobile visual search, visual vocabulary, boosting, compact visual descriptor

1. INTRODUCTION

Recent years have witnessed a great popularization of camera embedded mobile devices in our daily lives. With the ever growing wireless Internet services, there are great potentials for mobile landmark search with a wide range of applications. Generally speaking, most existing mobile landmark search systems follow a client-server architecture. A remote server maintains a large scale landmark database, where landmark photos are often tagged with GPS or human labeled locations. Near-duplicated search is applied with an inverted indexing mechanism, typically based on scalable bag-of-words models [1][8][9]. In online search, mobile users take a photo as his landmark query, which is transmitted to the remote server to identify its corresponding landmark. Consequently, its geographical location, photographing viewpoint, together with its related tourism and even commercial information, could be returned to the mobile user as query results.

In a typical scenario, the query photo transmission from the mobile device to the remote server is over a bandwidth-constrained wireless network. Undoubtedly, sending an entire image is time consuming, which would greatly degenerate the mobile search experiences. Arising from the ever growing mobile computing power, research efforts have been devoted to directly extracting landmark descriptors on a mobile device for the subsequent low cost query transmission. Comparing with previous works in compact local descriptors e.g. SURF [14] and PCA-SIFT [15], more recent works in [2][3][4][6] aim to achieve compression rates that are suitable for wireless transmission in mobile visual search scenarios.

Fig. 1. The developed mobile landmark search system.

The first group of works come from directly compressing the local visual descriptors. 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 describe each interest point using approximate 50 bits. The work in [6] compressed the SIFT descriptor with Karhunen-Loeve Transform, which yields approximate 2 bits per SIFT dimension. Tsai et al. [13] proposed to transmit the spatial layouts of interest points to improve the discriminability of CHoG descriptors.

The second group of works transmits the bag-of-features [2][3] instead of the original local descriptors, which gains much higher compression rates without much lost of discriminability. 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 to reduce the memory cost in a mobile device.

While compressing bag-of-features serves as a promising solution for low cost wireless transmission, previous works [2][3] still demand 2KB to 4KB visual descriptors per image. In an unstable wireless network, such cost would still delay the subsequent similarity ranking and results returning. In addition, two key issues remain ignored in the state-of-the-art compact visual descriptor design [2][3][4][6].

Location Insensitive Compression: The state-of-the-art works in compressive landmark descriptors [2][3][4][6]

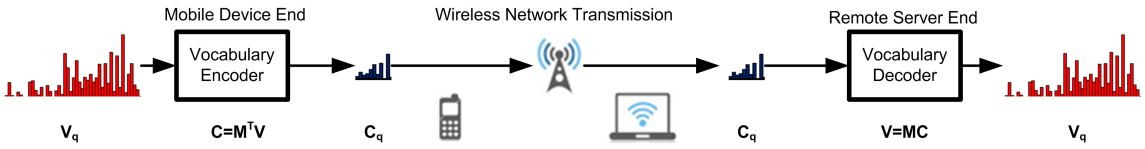

Fig. 2. The proposed vocabulary coding pipeline for low bit rate wireless transmission.

solely rely on the visual content of a query photo, regardless of its location context that can further improve the compression rate. On the contrary, more and more landmark photos are associated with ever increasing geographical tags like GPS or human labeled information [9][8][11]. Different geographical locations (e.g. different cities) often produce their specialized visual appearances, which should be incorporated into the design of effective compression strategy.

Vocabulary Generality: Targeting at real-world landmark search applications, while maintaining a single vocabulary at the worldwide scale cannot obtain satisfactory performance, current works prefer to maintain a vocabulary per city [9][10] to ensure high recognition rates. Hence in online search, when a mobile user enters a city, his mobile device has to load the corresponding vocabulary, which is inconvenient and extremely time consuming for mobile users.

In this paper, we present a location aware vocabulary coding scheme. In each city, we compress the vocabulary model via a vocabulary boosting strategy, which makes use of the most frequent landmark queries in the city to learn a compact yet discriminative vocabulary. Once a mobile user enters a city, the server transmits a downstream supervision (i.e., a compact BoW boosting list), which "teaches" the mobile device to transform the original vocabulary into a very compact one. With this technique, we have successfully developed a mobile landmark search prototype system, providing search services in Beijing, New York City, Lhasa, Singapore, and Florence (Snapshot in Figure 1). Quantitative comparisons with state-of-the-arts works in compact visual descriptor and signature [2][4][5] will be provided subsequently.

2. LOCATION AWARE VOCABULARY CODING

SVT Model for Scalable Search: Towards efficient landmark search in a million scale database, the Scalable Vocabulary Tree (SVT) [1] is well exploited in previous works [2][3][8][9]. SVT uses hierarchical k-means to partition local descriptors into quantized codewords. A H-depth SVT with B-branch produces $M = B^H$ codewords, and the scalable search typically settles $H = 6$ and $B = 10$ [1]. Given a query photo $\mathbf{I}_q$ with local descriptors $\mathbf{S}_q = [S^q_1, S^q_2, ..., S^q_J]$, SVT quantizes $\mathbf{S}_q$ by traversing in the vocabulary hierarchy to find the nearest codeword, which converts $\mathbf{S}_q$ to a BoW signature $\mathbf{V}_q = [V^q_1, V^q_2, ..., V^q_M]$. In search, desirable ranking aims to minimize the following ranking lost with respect to the ranking position $R(x)$ of each photo $\mathbf{I}_x$ in a n-photo database:

$$Lost_{Rank} = \sum_{x=1}^{N} R(x)\mathbf{W}_x \|\mathbf{V}_q, \mathbf{V}_x\|_{Cosine} \quad (1)$$

where TF-IDF weighting is calculated similar to its original form [16] in the document retrieval as:

$$\mathbf{W}_x = [\frac{n^x_1}{n^x} \times log(\frac{N}{N_{V_1}}), ..., \frac{n^x_M}{n^x} \times log(\frac{N}{N_{V_M}})] \quad (2)$$

n^x denotes the number of local descriptors in $\mathbf{I}_x$; $n^x_{V_i}$ denotes the number of local descriptors in $\mathbf{I}_x$ quantized into V_i; N denotes the total number of images in the database; N_{V_i} denotes the number of images containing V_i; $\frac{n^x_i}{n^x}$ serves as the term frequency of V_i in $\mathbf{I}_x$; and $log(\frac{N}{N_{V_i}})$ serves as the inverted document frequency of V_i in the database.

Modeling User Query Logs as Training Data: For a given city containing n landmark photos $[\mathbf{I}_1, \mathbf{I}_2, ..., \mathbf{I}_n]$, we collect a set of user query logs containing photos $[\mathbf{I}'_1, \mathbf{I}'_2, ..., \mathbf{I}'_{n_{sample}}]$, which outputs the following ranking list:

$$\begin{aligned} Query(\mathbf{I}'_1) &= [\mathbf{A}^1_1, \mathbf{A}^1_2, ..., \mathbf{A}^1_R] \\ ... &= ... \\ Query(\mathbf{I}'_{n_{sample}}) &= [\mathbf{A}^{n_{sample}}_1, \mathbf{A}^{n_{sample}}_2, ..., \mathbf{A}^{n_{sample}}_R] \end{aligned} \quad (3)$$

$\mathbf{A}^j_i$ is the ith returning of the jth query. We expect to still obtain $[\mathbf{A}^j_1, \mathbf{A}^j_2, ..., \mathbf{A}^j_R]$ for each jth query using a more compact vocabulary. Therefore, above queries and returning are treated as ground truth in our subsequent boosting training.

Location Aware Vocabulary Boosting: We deal with vocabulary coding as an AdaBoost based codeword selection: The weak learner is each single codeword, and learning is to minimize the ranking discriminability lost with a minimized coding length. We first define $[w_1, w_2, ..., w_{n_{sample}}]$ as an error weighting vector to the n_{sample} in the user query log, which measures the ranking consistency lost in the codeword selection. We then define the encoded vocabulary as $\mathbf{C} \in \mathbb{R}_K$, which is obtained from the original vocabulary $\mathbf{V} \in \mathbb{R}_M$ via $\mathbf{C} = \mathbf{M}^T\mathbf{V}$, where $\mathbf{M}_{M\times K}$ is a dimension reduction transform from $\mathbb{R}_M$ to $\mathbb{R}_K$. In the vocabulary boosting, $\mathbf{M}\mathbf{M}^T$ is a diagonal matrix, where each non-zero diagonal position defines a codeword selection/non-selection. At the tth iteration, we got the current $(t-1)$ non-zero diagonal elements in $\mathbf{M}\mathbf{M}^T$. To select the next tth discriminative codeword, we first estimate the ranking preservation of the current $\mathbf{M}\mathbf{M}^T$ as:

$$Lost(\mathbf{I}'_i) = w^{t-1}_i \sum_{r=1}^{R} R(\mathbf{A}^1_r)\mathbf{W}_{\mathbf{A}^1_r} \|\mathbf{M}^{t-1}\mathbf{C}_{\mathbf{I}'_i}, \mathbf{V}_{\mathbf{A}^i_r}\|_{Cosine} \quad (4)$$

Algorithm 1: Location Aware Vocabulary Boosting

Input: bag-of-words signatures $\mathbf{V} = \{\mathbf{V}_i\}_{i=1}^{n}$; user query logs $\{Query(\mathbf{I}'_r)\}_{r=1}^{R}$; Boosting threshold τ; error weighting vector $[w_1, w_2, ..., w_{n_{sample}}]$; and boosting iteration $t = 0$.
Pre-Computing: Calculate $Lost_{Rank}$ in each city using Equation 5; Calculate $\sum_{i=1}^{n_{sample}} w_i^t$
while $\{\sum_{i=1}^{n_{sample}} w_i^t \leq \tau\}$ **do**
    Lost Estimation: Calculate $Lost_{Rank}$ by Equation 5.
    Codeword Selection: Select C_t by Equation 6.
    Error Weighting: Update $[w_1, ..., w_{n_{sample}}]$ by Equation 7;
    Transformation Renew: Update $\mathbf{M}^{t-1}$ by Equation 8.
    $t++$;
end
Output: The compressed codebook $\mathbf{C}_{region} = \mathbf{M}^T\mathbf{V}$.

where $i \in [1, n_{sample}]$; $R(\mathbf{A}_r^i)$ is the current position of the originally ith returning of query $\mathbf{I}'_i$; $[w_1^{t-1}, w_2^{t-1}, ..., w_{n_{sample}}^{t-1}]$ is the *(t-1)*th error weighting, measuring the ranking lost of the jth query ($j \in [1, n_{sample}]$). Then, the overall ranking lost is:

$$Lost_{Rank} = \sum_{i=1}^{n_{sample}} w_i^{t-1} \sum_{r=1}^{R} R(\mathbf{A}_r^i)\mathbf{W}_{\mathbf{A}_r^i} \|\mathbf{M}^{t-1}\mathbf{C}_{\mathbf{I}'_i}, \mathbf{V}_{\mathbf{A}_r^i}\|_{Cosine} \tag{5}$$

the best new codeword C_t is selected by minimizing:

$$C_t = \arg\min_j \sum_{i=1}^{n_{sample}} w_i^{t-1} \sum_{r=1}^{R} R(\mathbf{A}_r^i)\mathbf{W}_{\mathbf{A}_r^i} \times \|\mathbf{V}_{\mathbf{A}_r^i}, [\mathbf{M}^{t-1} + [0, ..., pos(j), .., 0]_M [0, ..., pos(t), .., 0]_K^T]\mathbf{C}_{\mathbf{I}'_i}\|_{Cosine} \tag{6}$$

where $[0, ..., pos(j), .., 0]_M$ is a $M \times 1$ selection vector, which selects the jth column into the linear projection; And the vector $[0, ..., pos(t), .., 0]_K$ is a $K \times 1$ position vector, which maps v_j into the new codeword C_t. Subsequently, we update the error weighting of each w_i^{t-1} as:

$$w_i^t = \sum_{r=1}^{R} R(\mathbf{A}_r^i)\mathbf{W}_{\mathbf{A}_r^i} \|\mathbf{V}_{\mathbf{A}_r^i}, [\mathbf{M}^{t-1} + [0, ..., pos(j), .., 0]_M [0, ..., pos(t), .., 0]_K^T]\mathbf{C}_{\mathbf{I}'_i}\|_{Cosine} \tag{7}$$

Also, the $\mathbf{M}$ at the tth round is updated as follows:

$$\mathbf{M}^t = \mathbf{M}^{t-1} + [0, ..., pos(j), .., 0]_M [0, ..., pos(t), .., 0]_K^T \tag{8}$$

The codebook boosting is stopped at $\sum_{i=1}^{n_{sample}} w_i^t \leq \tau$. We summarize our Vocabulary Boosting in Algorithm 1.

3. IMPLEMENTATIONS AND RESULTS

User Interface Design: Figure 1 shows the user interface of our mobile landmark search system, which is currently deployed on a HTC DESIRE G7 intelligent mobile phone.

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. Figure 3 shows the geographical photo distribution in Beijing.

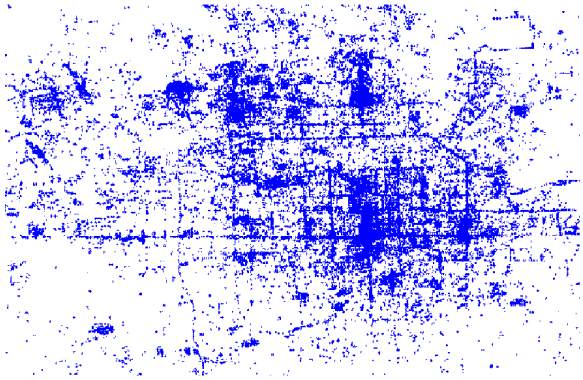

Fig. 3. The geographical distribution of Flickr and Panoramio photos in Beijing. This GPS distribution can reveal users' photographing behaviors during touring this city.

Simulating User Query Logs: From the geographical map of each city, we choose 30 most dense regions and 30 randomly selected regions. Since manually identifying all related photos of a landmark is intensive, for each of these 60 regions, we ask volunteers to manually identify one or more dominant views. All near-duplicated landmark photos are labeled in its current and nearby regions. Then, we sample 5 images from each region to form the ground truth. It generates in total 300 user query logs for each city.

Parameters and Evaluation: From Beijing landmark photo collection, we extract both SIFT [7] and CHoG [4] features from each photo. Then, we build a Scalable Vocabulary Tree [1] to generate the initial Vocabulary $\mathbf{V}$, which generates a bag-of-words signature $\mathbf{V}_i$ for each database photo $\mathbf{I}_i$. We use the vocabulary generated in Beijing to do search in five cities, for each of which the boosting is carried out to build the $\mathbf{M}$ transformation. We denote the hierarchical level as H and the branching factor as B. In a typical settlement, we have $H = 6$ and $B = 10$, producing approximate 100,000 codewords. We use Mean Average Precision at N ($MAP@N$) to evaluate our system performance, which reveals its position-sensitive ranking precision in the top N positions.

Baselines: (1) *Initial bag-of-words:* Transmitting the entire BoW has the lowest compression rate. However, it provides the upper bound in MAP. (2) *Word Frequency Compression:* As the most straightforward scheme, we retain the codewords with the top 20% highest IDF as the vocabulary compression result. (3) *Aggregating Local Descriptors [5]:* The work in [5] leverages aggregate quantization to obtain compact signature. Its input is also the bag-of-words signature produced by the initial vocabulary $\mathbf{V}$. (4) *Tree Histogram Coding [2]:* Chen et al. used residual coding scheme over the BoW histogram, which Is the most related work to ours.

Rate Distortion Analysis: We give the rate distortion analysis in comparisons with state-of-the-art works in [2][4]

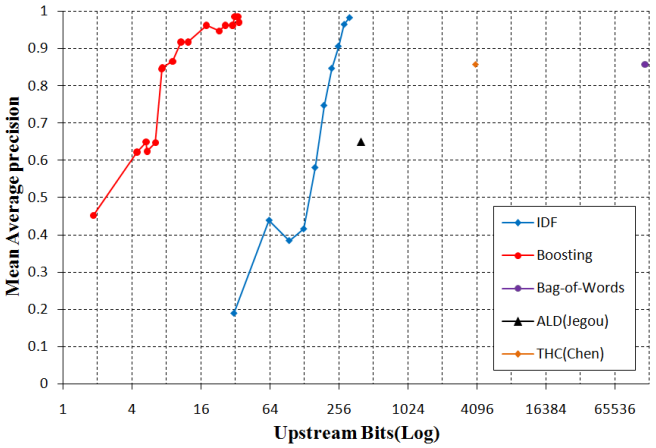

Fig. 4. Compression rate and ranking distortion comparing with [2][4][5] using our ground truth query set.

[5]. Figure 4 shows that we achieve the highest compression rates with equalized distortion (horizontal viewpoint), or maintain the highest MAP ranking performance with equalized compression rates (vertical viewpoint).

Case Study: We discovered that some empirical queries in our system happen to be taken at night. And some queries occur in different scales (from either nearby views or distant views). There is also a common problem that queries are with blurs (very common cases). We also selected some suboptimal queries with partial occlusions (objects or persons), as well as photos of partial landmark views. Figure 5 shows that our vocabulary coding can still well preserve the ranking precision, with comparisons to the Baselines (1)(4).

4. CONCLUSIONS

We have presented a novel location aware vocabulary coding scheme, which embeds location cues to boost a compact vocabulary subset for effective and efficient landmark search. Our prototype system has demonstrated superior performance over state-of-the-art works in mobile visual descriptor and compact image signatures. Nowadays consumers increasingly seek relevancy in mobile technology. Mobile users are quick to adopt location-based technology. We envision that the location sensitive compression could bring great benefits in real world mobile photo search for e-guide, such as taking a photo and sending it to a cloud, and getting back a guild service on landmark description, campus direction, museum introduction, and so on.

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.

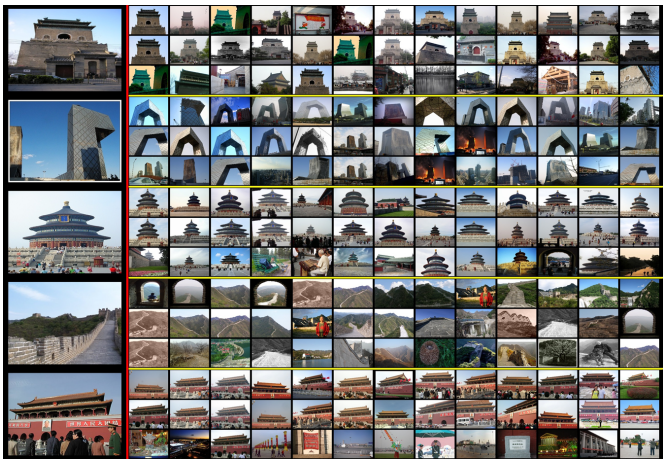

Fig. 5. Case study of illumination changes, scale changes, blurred photographing, occlusions, and partial landmark queries. Each photo on the left is the query, each line of returning results corresponds to an approach. Top: Vocabulary Boosting; Middle: Original BoW feature or Tree Histogram Coding [2]; Bottom: IDF Thresholding (top 20%).

6. REFERENCES

[1] Nister D. and Stewenius H. Scalable recognition with a vocabulary tree. *CVPR*. 2006.
[2] Chen D., Tsai S., and Chandrasekhar V. Tree histogram coding for mobile image matching. *DCC*. 2009.
[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.
[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.
[5] Jegou H., Douze M., Schmid C., Perez P. Aggregating local descriptors into a compact image representation. *CVPR*. 2010.
[6] Chandrasekhar V., Takacs G., Chen D., et. al. and Girod B. Transform coding of image feature descriptors. *VCIP*. 2009.
[7] Lowe D. G. Distinctive image features from scale-invariant keypoints. *IJCV*. 2004.
[8] Irschara A., Zach C., Frahm J., Bischof H. From SFM point clouds to fast location recognition. *CVPR*. 2009.
[9] Schindler G. and Brown M. City-scale location recognition. *CVPR*. 2007.
[10] Ji R., Xie X., Yao H., and Ma W.-Y. Hierarchical optimization of visual vocabulary for effective and transferable retrieval. *CVPR*. 2009.
[11] Crandall D., Backstrom L., Huttenlocher D., and Kleinberg J. Mapping the world's photos. *WWW*. 2009.
[12] Ji R., Xie X., Yao H., Ma W.-Y., and Wu Y. Vocabulary tree incremental indexing for scalable scene recognition. *ICME*. 2008.
[13] Tsai S., Chen D., Takacs G., Chandrasekhar V. Location coding for mobile image retrieval. *MobileMedia*. 2010.
[14] Bay H., Tuytelaars T., and Van Gool L. SURF: Speeded up robust features. *ECCV*. 2006.
[15] Ke Y. and Sukthankar R. PCA-SIFT: A more distinctive representation for local image descriptors. *CVPR*. 2004.
[16] Salton G. and Buckley C. Term-weighting approaches in automatic text retrieval. *Info. Proc. and Management*. 1998.