Learning Compact Visual Descriptor for Low Bit Rate Mobile Landmark Search

Rongrong Ji[⋆†] **Ling-Yu Duan**[⋆] **Jie Chen**[⋆] **Hongxun Yao**[†] **Tiejun Huang**[⋆] **Wen Gao**[⋆†]
[⋆]Institute of Digital Media, Peking University, Beijing, 100871, China
[†]Visual Intelligence Laboratory, Harbin Institute of Technology, Heilongjiang, 150001, China
{lingyu,cjie,tjhuang,wgao}@pku.edu.cn {rrji,yhx}@vilab.hit.edu.cn

Abstract

In this paper, we propose to extract a compact yet discriminative visual descriptor directly on the mobile device, which tackles the wireless query transmission latency in mobile landmark search. This descriptor originates from offline learning the location contexts of geo-tagged Web photos from both Flickr and Panoramio with two phrases: First, we segment the landmark photo collections into discrete geographical regions using a Gaussian Mixture Model [Stauffer *et al.*, 2000]. Second, a ranking sensitive vocabulary boosting is introduced to learn a compact codebook within each region. To tackle the locally optimal descriptor learning caused by imprecise geographical segmentation, we further iterate above phrases incorporating the feedback of an "entropy" based descriptor compactness into a prior distribution to constrain the Gaussian mixture modeling. Consequently, when entering a specific geographical region, the codebook in the mobile device is downstream adapted, which ensures efficient extraction of compact descriptors, its low bit rate transmission, as well as promising discrimination ability. We descriptors to both HTC and iPhone mobile phones, testing landmark search over one million images in typical areas like Beijing, New York, and Barcelona, *etc.* Our descriptor outperforms alternative compact descriptors [Chen *et al.*, 2009][Chen *et al.*, 2010][Chandrasekhar *et al.*, 2009a][Chandrasekhar *et al.*, 2009b] with a large margin.

1 Introduction

With the popularization of mobile embedded cameras, there is a great potential for mobile landmark search with a wide range of applications, such as location recognition, scene retrieval, and photographing recommendation. In general, most existing mobile landmark search systems are deployed under a client-server architecture: In the server end, a visual search system is maintained, typically based on scalable BoW models [Nister *et al.*, 2006][Irschara *et al.*, 2009][Schindler *et al.*, 2007], where landmark photos, as well as their geographically tags like GPS, are inverted indexed into a visual vocabulary. In online search, a landmark query is sent through the wireless network to the server, where near-duplicated search is conducted to identify its best matched landmark, and subsequently returns its geographical location and touristic information to the mobile user.

In a typical scenario, the query photo transmission is over a bandwidth-constrained wireless network. With the ever growing computational power in the mobile devices, while sending the entire query is often unnecessary and time consuming, we propose to perform the visual descriptor extraction directly on the mobile end. In this scenario, the expected visual descriptor should be compact, discriminative, and meanwhile efficient for extraction to tackle the wireless query transmission latency, which also receives dedicated efforts in industry standards like MPEG [Yuri *et al.*, 2010].

Towards low bit rate mobile landmark search, previous local descriptors, e.g. SIFT [Lowe 2004], SURF [Bay *et al.*, 2006], and PCA-SIFT [Ke *et al.*, 2004] cannot work well due to their descriptor lengths. We also argue that recent works in compact visual descriptors [Chen *et al.*, 2009][Chen *et al.*, 2010][Chandrasekhar *et al.*, 2009a][Chandrasekhar *et al.*, 2009b] are still not compact enough towards zero-latency wireless transmission, which is quantitatively proven later. This paper proposes to achieve descriptor compactness through "*contextual learning*", with additional attentions of the mobile end extraction efficiency.

Our Contribution: We explore this "*contextual learning*" in combination with the visual statistics in each specific landmark region to learn a compact descriptor from geo-tagged reference photo collections: First, we propose a geographical segmentation scheme based on Gaussian Mixture Model [Stauffer *et al.*, 2000]; Second, we introduce a vocabulary boosting scheme to learn a compact descriptor in each specific region, which simulates a set of landmark queries from this region and learns a compact codebook to maintain the ranking precision from an original visual vocabulary [Sivic *et al.*, 2003][Nister *et al.*, 2006]. With this codebook, a compact BoW descriptor is generated for a given query. However, due to imprecise segmentation, learning compact descriptors respectively in each individual region cannot guarantee a global optimum (as discussed in Section 3). Hence, we propose to further iterate the content aware geographical segmentation and the vocabulary boosting to reinforce each other. Figure 2 shows the mobile visual landmark search system, embedded

Figure 1: The developed mobile visual landmark search system, which embeds the compact visual descriptor extraction in HTC Desire G7 mobile phone, testing in typical areas like Beijing, New York, and Barcelona, *etc.*

with contextual learning based compact visual descriptors.

Application Scenarios: Once a mobile user enters a given region, the server transmits a downstream supervision (i.e. a compact codeword boosting vector) to "*teach*" the mobile device by linearly projecting the original high-dim vocabulary into a compact codebook via this boosting vector. Then, given a query, instead of high-dimensional codeword histogram, an extremely compact histogram is redirected to transmit via 3G or WLAN wireless network.

Paper Outlines: We review related work in visual vocabulary and compact visual descriptors in Section 2. Then, our contextual learning based descriptor extraction is given in Section 3. Section 4 introduces our descriptor implementation in mobile visual search system, covers typical areas like Beijing, New York City, and Barcelona (see snapshot in Figure 1), with quantitative comparisons with the state-of-the-art compact visual descriptors [Chen *et al.*, 2009][Chen *et al.*, 2010][Chandrasekhar *et al.*, 2009a][Chandrasekhar *et al.*, 2009b].

2 Related Work

Visual Vocabulary Construction: The standard approach to building visual vocabulary usually resorts to unsupervised vector quantization such as K-means [Sivic *et al.*, 2003], which subdivides local feature space into codeword regions. An image is then represented as a BoW histogram, where each bin counts how many local features of this image fall into its corresponding codeword. In recent years, there are many vector quantization based vocabularies, such as Vocabulary Tree [Nister *et al.*, 2006], Approximate K-means [Philbin *et al.*, 2007], Hamming Embedding [Jegou *et al.*, 2008], and [Jurie *et al.*, 2005][Jiang *et al.*, 2007][Philbin *et al.*, 2007][Jegou *et al.*, 2010] *et al.* Hashing based approach is another solution, such as Locality Sensitive Hashing and its kernelized version [Kulis *et al.*, 2009]. The works in [Jiang *et al.*, 2007][Jegou *et al.*, 2008][Philbin *et al.*, 2007][Gemert *et al.*, 2009] also handles codeword uncertainty and ambiguity, *e.g.* Hamming Embedding [Jegou *et al.*, 2008], Soft Assignments [Philbin *et al.*, 2007], and kernelized codebook [Gemert *et al.*, 2009]. Recent works in [Moosmann *et al.*, 2006][Mairal *et al.*, 2008][Lazebnik *et al.*, 2009] made use of semantics or category labels to supervise the vocabulary construction.

Compact Visual Descriptors for Mobile Visual Search: Comparing with previous works in compact local descriptors, e.g. SURF [Bay *et al.*, 2006] and PCA-SIFT [Ke *et al.*, 2004], more recent works [Chen *et al.*, 2009][Chen *et al.*, 2010][Chandrasekhar *et al.*, 2009a][Chandrasekhar *et al.*, 2009b] aimed to achieve desirable compression rates that suffice for 3G wireless transmission in mobile visual search scenarios.

The first group comes from direct compression of local visual descriptors. For instance, Chandrasekhar *et al.* proposed a Compressed Histogram of Gradient (CHoG) [Chandrasekhar *et al.*, 2009a] for compressive local feature description, which adopted both Huffman Tree and Gagie Tree to describe each interest point using approximate 50 bits. The work in [Chandrasekhar *et al.*, 2009b] compressed the SIFT descriptor with Karhunen-Loeve Transform, which yielded approximate 2 bits per SIFT dimension. Tsai *et al.* [Tsai *et al.*, 2010] proposed to transmit the spatial layouts of interest points to improve the discriminability of CHoG descriptors. Considering the the successive but order insensitive delivery among local features, recent work in [Chen *et al.*, 2011] also proposed to sort the local features the inter-feature compression in addition to the intra-feature compression.

The second group transmits the BoW [Chen *et al.*, 2009][Chen *et al.*, 2010] instead of the original local descriptors to gain much higher compression rate without serious loss of discriminability. Chen *et al.* [Chen *et al.*, 2009] proposed to compress the sparse bag-of-features 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 [Chen *et al.*, 2010] further compressed the inverted indices of vocabulary tree [Nister *et al.*, 2006] with arithmetic coding to reduce the memory cost in a mobile device. Recent work in [Ji *et al.*, 2011] also proposed to compress the visual vocabulary within the entire city for city-scale mobile landmark search.

3 Learning Compact Landmark Descriptor

Problem Formulation: We denote scalars as italic letters, e.g. v; denote vectors as bold italic letters, e.g. $\mathbf{v}$; denote instance spaces for n instances as $\mathbb{R}_n$; and denote the inner product between $\mathbf{u}$ and $\mathbf{v}$ as $\mathbf{u} \cdot \mathbf{v} = \sum_{i=1}^{n} u_i v_i$.

Given database images $\mathbf{I} = \{I_i\}_{i=1}^{n}$, we offline extract n Bag-of-Words histograms [Nister *et al.*, 2006][Sivic *et al.*, 2003] $\mathbf{V} = \{V_i\}_{i=1}^{n}$, which are typically high-dimensional, say 0.1-1 million in state-of-the-art settlings [Nister *et al.*, 2006]. In addition, all images are bounded with corresponding GPS coordinates as $\mathbf{G} = \{Lat_i, Long_i\}_{i=1}^{n}$.

Learning Goal: We aim to (1) learn a geographical segmentation $\mathbf{S} = \{S_j\}_{j=1}^{m}$ to partition $\mathbf{I} = \{I_i\}_{i=1}^{n}$ into m regions, which attempts to represent the local context to achieve descriptor compactness to an extreme. (2) learn a codebook $\mathbf{U}_j \in \mathbb{R}_k$ for compact descriptor extraction in each S_j from $\mathbf{V} \in \mathbb{R}_n$ such that $k \ll n$, which is online updated into the mobile device once the mobile user enters S_j.

Chicken and Egg Problem: On one hand, we expect the $\mathbf{U}_j$ is as compact as possible in each S_j. On the other hand, under such circumstance, the compactness depends on how properly an image subset in $\mathbf{I}$ is segmented into S_j. However, such segmentation is naturally imprecise, especially in

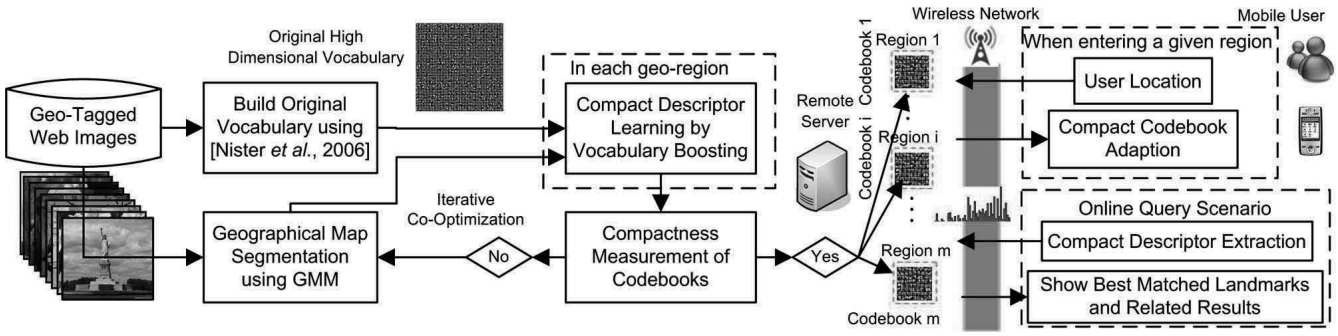

Figure 2: The proposed contextual learning based compact visual landmark descriptor extraction framework towards a low bit rate mobile landmark search, which embeds descriptor extraction into the mobile end.

the context of learning compact visual descriptors. While we may learn an optimal descriptor in each region, the overall compactness of all regions may not be guaranteed well. In other words, the optimization towards descriptor compactness is *local* in each region, rather than among all the regions over the entire image database.

Ideally, we aim to learn both the optimal region segmentation and more compact description in each region to minimize:

$$Cost = \sum_{j=1}^{m} \sum_{i=1}^{n'} |U_i| \quad s.t. \quad \forall j \in m \;\; Loss(P_{S_j}) \leq T \quad (1)$$

where $|U_i|$ denotes the descriptor length of the *i*th sampled query image (in total n') falling into region S_j; the constraints denote the retrieval precision loss ($Loss(P_{S_j})$) in each region, which would be revisited in Section 3.2. Obviously, we cannot carry out both geographical segmentation $\mathbf{S} = \{S_j\}_{j=1}^{m}$ and descriptor learning $\mathbf{U}_j \in \mathbb{R}_k$ in each S_j simultaneously. Hence, we expect an iterative learning to optimize Equation 1 in the entire city.

3.1 Geographical Map Segmentation

We adopt the Gaussian Mixture Model (GMM) [Stauffer *et al.*, 2000] to segment $\mathbf{I}$ into $\mathbf{S}$. We assume that the geographical photo distribution is drawn from m landmark regions, and denote the *i*th component as w_i, with mean vector μ_i. We then regard photos belonging to the *i*th component as generated from the *i*th Gaussian with mean μ_i and covariance matrix $\sum_i$, followed a normalized distribution $N(\mu_i, \sum_i)$.

Therefore, assigning each photo x into the *i*th region is to infer its Bayesian posterior probability:

$$p(y = i|x) = \frac{p(x|y = i)P(y = i)}{p(x)} \quad (2)$$

where $p(x|y = i)$ is the probability that the region label y of photo x belongs to the *i*th component, following a normalized distribution:

$$p(x|y = i) = \frac{1}{(2\pi)^{\frac{m}{2}} ||\sum_i||^{\frac{1}{2}}} exp\left[-\frac{1}{2}(x - \mu_i)^T(x - \mu_i)\right] \quad (3)$$

As neither the component parameters nor the region assignments is known, we adopt an expectation maximization to perform segmentation: First, we estimate Gaussian Mixture Model at the *t*th iteration (denoted by (t)) as:

$$\lambda_t = \left\{\mu_1(t), ...\mu_m(t), \sum_1(t), ... \sum_m(t), P_1(t), ...P_m(t)\right\} \quad (4)$$

The **Expectation Step** computes the "expected" segmentation label for each database image x as:

$$\begin{aligned} p(y = i|x, \lambda_t) &= \frac{p(x|y = i, \lambda_t)p(y = i|\lambda_t)}{p(x|\lambda_t)} \\ &= \frac{p\left(x|y = i, \mu_i(t), \sum_i(t)\right) P_i(t)}{\sum_{j=1}^{m} p\left(x|y = j, \mu_j(t), \sum_j(t)\right) P_j(t)} \end{aligned} \quad (5)$$

The **Maximization Step** computes the maximal likelihood of each component, given the segmentation membership of x:

$$\mu_i(t + 1) = \frac{\sum_k p(y_k = i|x_k, \lambda_t)x_k}{\sum_k p(y_k = i|x_k, \lambda_t)} \quad (6)$$

$$\sum_i(t + 1) = \frac{\sum_k p(y_k = i|x_k, \lambda_t)[x_k - \mu_i(t + 1)][x_k - \mu_i(t + 1)]^T}{\sum_k p(y_k = i|x_k, \lambda_t)} \quad (7)$$

The probability of the *i*th component P_i specifies its priori in geographical segmentation, which is updated as follows:

$$P_i(t + 1) = \frac{\sum_k p(y_k = i|x_k, \lambda_t)}{|\{I_k|k \in [1, n],\ y_k = S_m\}|} \quad (8)$$

We revisit P_i in our iterative co-optimization (Section 3.3) to learn geographical segmentation and compact descriptors in a joint manner.

3.2 Descriptor Learning via Vocabulary Boosting

SVT Model for Scalable Search: Towards efficient landmark search in a million scale database, the Scalable Vocabulary Tree (SVT) [Nister *et al.*, 2006] is well exploited in previous works [Chen *et al.*, 2009][Chen *et al.*, 2010][Irschara *et*

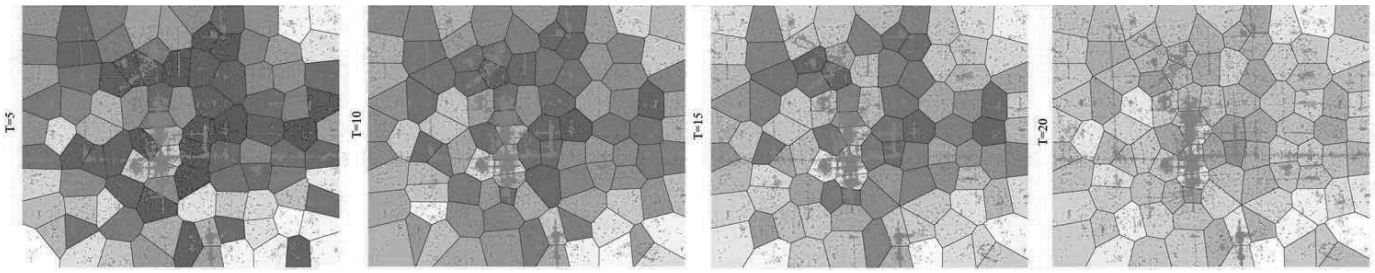

Figure 3: The geographical visualization of the descriptor compactness in Beijing city through iterative co-optimization ($T = 1\ to\ 20$). We normalize the min vs. max ratio of descriptor lengths and map the ratio to the saturation of red color. The green points denote the distribution of geo-tagged photos. In general, less saturated map corresponds to more optimal descriptors.

al., 2009][Schindler *et al.*, 2007]. SVT uses hierarchical K-means to partition local descriptors into quantized codewords. An H-depth SVT with B-branch produces $M = B^H$ codewords, and the scalable search typically settles $H = 5$ and $B = 10$ [Nister *et al.*, 2006]. Given a query photo I_q with J local descriptors $\mathbf{L}(q) = [L_1(q), ..., L_J(q)]$, SVT quantizes $\mathbf{L}(q)$ by traversing in the vocabulary hierarchy to find out the nearest codeword, which converts $\mathbf{L}(q)$ to a BoW histogram $\mathbf{V}(q) = [V_1(q), ..., V_M(q)]$. In search, the optimal ranking is supposed to minimize the following loss function with respect to the ranking position $R(x)$ of each I_x (BoW feature $\mathbf{V}(x)$):

$$Loss_{Rank} = \sum_{x=1}^{n} R(x)\mathbf{W}_x||\mathbf{V}(q), \mathbf{V}(x)||_{L2} \tag{9}$$

where TF-IDF weighting is calculated in a way similar to its original form [Salton *et al.*, 1988] in document retrieval:

$$\mathbf{W}_x = [\frac{n_1^x}{n^x} \times log(\frac{n}{n_{V_1}}), ..., \frac{n_M^x}{n^x} \times log(\frac{n}{n_{V_M}})] \tag{10}$$

where n^x denotes the number of local descriptors in I_x; $n_{V_i}(x)$ denotes the number of local descriptors in 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; $\frac{n_i^x}{n^x}$ serves as the term frequency of V_i in I_x; and $log(\frac{n}{n_{V_i}})$ serves as the inverted document frequency of V_i in the database.

Simulating User Query for Training: For a given region containing n' landmark photos $[I_1, I_2, ..., I_{n'}]$, we randomly sample a set of geo-tagged photos $[I'_1, I'_2, ..., I'_{n_{sample}}]$ as pseudo queries, which output the following ranking list:

$$\begin{aligned} Query(I'_1) &= [A_1^1, ..., A_R^1] \\ &... \\ Query(I'_{n_{sample}}) &= [A_1^{n_{sample}}, ..., A_R^{n_{sample}}] \end{aligned} \tag{11}$$

where A_i^j is the *i*th returning of the *j*th query. We expect to maintain the original ranking list $[A_1^j, A_2^j, ..., A_R^j]$ for the *j*th query using a more compact vocabulary. Therefore, above queries and results are dealt with as ground truth in our subsequent boosting training.

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

$$\begin{aligned} &Loss(I'_i) = \\ &w_i^{t-1} \sum_{r=1}^{R} R(A_r^1)\mathbf{W}_{A_r^1}||\mathbf{M}^{t-1}\mathbf{U}_j(I'_i), \mathbf{V}(A_r^i)||_{L2} \end{aligned} \tag{12}$$

where $i \in [1, n_{sample}]$; $R(A_r^i)$ is the returning position of the originally *r*th returning for query I'_i; $[w_1^{t-1}, ..., w_{n_{sample}}^{t-1}]$ is the *(t-1)*th error weighting, measuring the loss of the *j*th query. Then, the overall loss in ranking is:

$$\begin{aligned} &Loss_{Rank} = \\ &\sum_{i=1}^{n_{sample}} w_i^{t-1} \sum_{r=i}^{R} R(A_r^i)\mathbf{W}_{A_r^i}||\mathbf{M}^{t-1}\mathbf{U}_j(I'_i), \mathbf{V}(A_r^i)||_{L2} \end{aligned} \tag{13}$$

The best new codeword U_t is selected by minimizing:

$$\begin{aligned} &U_t = \arg\min_j \\ &\sum_{i=1}^{n_{sample}} w_i^{t-1} \sum_{r=1}^{R} R(A_r^i)\mathbf{W}_{A_r^i} \times ||\mathbf{V}(A_r^i), [\mathbf{M}^{t-1} + \\ &[0, ..., pos(j), .., 0]_M [0, ..., pos(t), .., 0]_K^T]\mathbf{U}_j(I'_i)||_{L2} \end{aligned} \tag{14}$$

where $[0, ..., pos(j), .., 0]_M$ is an $M \times 1$ selection vector to select the *j*th column to the linear projection; and

Algorithm 1: Location Aware Vocabulary Boosting

Input: Bag-of-Words $\mathbf{V} = \{\mathbf{V}(i)\}_{i=1}^{n'}$ for region S_j; simulated query logs $\{I'_i\}_{i=1}^{n_{Sample}}$; Boosting threshold τ; initial error weighting vector $[w_1, w_2, ..., w_{n_{sample}}]$; and boosting iteration $t = 0$.
Pre-Computing: Calculate $Loss_{Rank}$ in each region using Equation 13; Calculate $\sum_{i=1}^{n_{sample}} w_i^t$.
while $\{\sum_{i=1}^{n_{sample}} w_i^t \leq \tau\}$ **do**
  Loss Estimation: Calculate $Loss_{Rank}$ by Equation 13.
  Codeword Selection: Select U_t by Equation 14.
  Error Weighting: Update $[w_1, ..., w_{n_{sample}}]$ by Equation 15;
  Transformation Renew: Update $\mathbf{M}^{t-1}$ by Equation 16.
  $t++$;
end
Output: Compressed codebook $\mathbf{U}_j = \mathbf{M}^T\mathbf{V}$ for region S_j.

$[0, ..., pos(t), .., 0]_K$ is a $K \times 1$ position vector to map V_j to the new codeword U_t. We then update the error weighting of each w_i^{t-1}:

$$w_i^t = \sum_{r=1}^{R} R(A_r^i)\mathbf{W}_{A_r^i}||\mathbf{V}(A_r^i), [\mathbf{M}^{t-1} + [0, ..., pos(j), .., 0]_M[0, ..., pos(t), .., 0]_K^T]\mathbf{U}_j(I'_i)||_{L2} \quad (15)$$

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

$$\mathbf{M}^t = \mathbf{M}^{t-1} + [0, ..., pos(j), .., 0]_M[0, ..., pos(t), .., 0]_K^T \quad (16)$$

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

3.3 Iterative Co-Optimization

Since there is a tradeoff between the downstream vocabulary adaption and the upstream descriptor delivery, we aim to subdivide the geographical regions, which would yield more compact descriptors and meanwhile aim to merge nearby regions towards less compact descriptors. We fulfill these two joint goals by iterative co-optimization, through the process of geographical segmentation as well as descriptors learnings.

We estimate the necessity of using longer or shorter descriptor length of each region with an uncertainty measurement (like "entropy"), which is dealt with as a feedback to the prior probability P_i in Equation 2 (for the ith region) to refine geographical segmentation at the Tth iteration:

$$P_i = -log\left(|\mathbf{U}_i|/|\mathbf{U}_{max}|\right) \quad (17)$$

Figure 3 further visualizes the city-scale descriptor minimization procedure with the segmentation-learning iteration ($T = 1\ to\ 20$). It is intuitive to find out that the overall descriptor length minimization is gradually reduced. In other words, when the mobile users travel in this city and visit multiple landmarks, or multiple travelers visit multiple landmark regions, the overall downstream descriptor adaption and upstream wireless query transmission, is minimized in a more global manner.

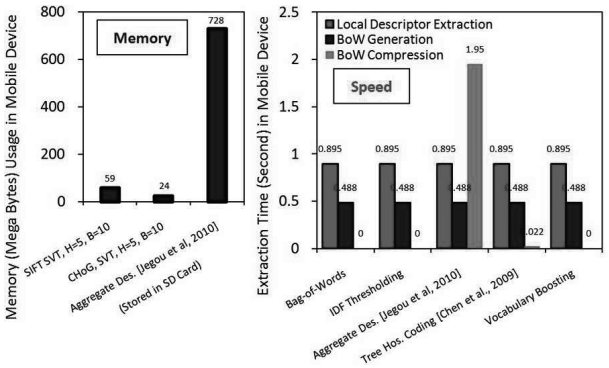

Figure 4: Memory and time cost comparison between our proposed compact descriptor and the state-of-the-art compact descriptors [Jegou *et al.* 2010][Chandrasekhar *et al.* 2009] in the HTC DESIRE G7 mobile phone.

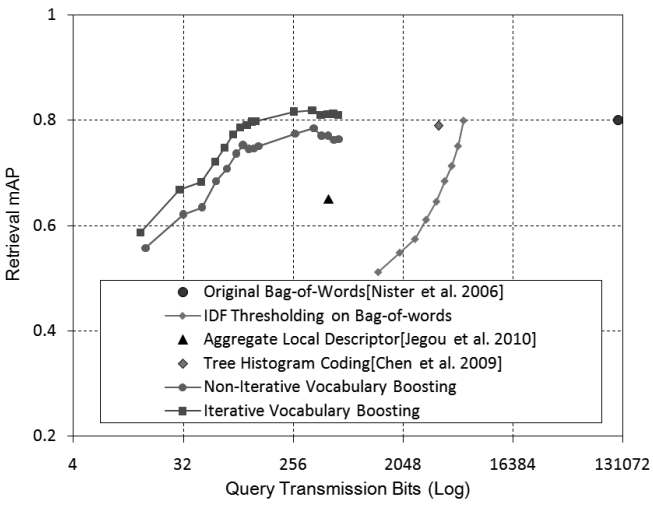

Figure 5: Compression rate and ranking distortion of our descriptor learning comparing with state-of-the-arts.

4 Implementations and Quantitative Comparison Results

Data Collection and Ground Truth Labeling: We collect over one million geographical tagged photos from both Flickr and Panoramio photo sharing websites, which cover typical areas including Beijing, New York City, and Barcelona. From the geographical map of each city, we choose 30 most dense regions and 30 random 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. Then, all near-duplicated landmark photos to a given view are labeled in its belonging and nearby regions. We then sample 5 images from each region as the query, which forms 300 queries in each city.

Parameters and Evaluations: We extract SIFT [Lowe 2004] features from each photo. We build a Scalable Vocabulary Tree [Nister *et al.*, 2006] to generate the initial vocabulary $\mathbf{V}$, which outputs a Bag-of-Words signature $\mathbf{V}(i)$ for

each database photo $\mathbf{I}_i$. We use the identical vocabulary generated in Beijing to conduct search in other cities. For each region in a city, the boosting is conducted to offline learn $\mathbf{M}$. We denote the hierarchical level as H and the branching factor as B. In a typical settlement, we have $H = 5$ and $B = 10$, producing approximate 0.1 million words. We use mean Average Precision (mAP) to evaluate performance, which reveals its position-sensitive ranking precision in top *10* positions.

Baselines: (1) *Original Bag-of-Words*: Transmitting the entire BoW has the lowest compression rate. However, it provides the upper bound in mAP. (2) *IDF Thresholding*: As a straightforward scheme, we only transmit the IDs of codewords with the highest IDF values (Figure 5 tests $20\% - 100\%$) as an alternative solution for vocabulary Compression. (3) *Aggregating Local Descriptors* [Jegou *et al.*, 2010]: The work in [Jegou *et al.*, 2010] adopted aggregate quantization to obtain compact signature. Its output is also a compressed codeword histogram produced by an initial vocabulary $\mathbf{V}$. (4) *Tree Histogram Coding* [Chen *et al.*, 2009]: Chen *et al.* used residual coding to compress the BoW histogram, which is the most related work. (5) *Without Co-Optimization*: To quantize our iterative co-optimization, we degenerate our approach without iterating between geographical segmentation and descriptor learning.

Efficiency Evaluation: We deployed our compact descriptor extraction into both HTC DESIRE G7 and iPhone4 (the upstream query transmission rate is given in Figure 5). Figure 4 gives a typical memory and time requirement in separate steps when embedding the complete descriptor extraction into the mobile phone.

Rate Distortion Analysis: We compare our rate distortion with state-of-the-art works [Nister *et al.*, 2006][Chen *et al.*, 2009][Chandrasekhar *et al.*, 2009a][Jegou *et al.*, 2010] in Figure 5. We achieve the highest compression rate with equalized distortions (horizontal view), or in other words, maintain the highest search mAP with equalized compression rates (vertical view).

Insights into Compact Descriptors: *Descriptor Robustness and Matching Locations*: We empirically collect quite a few real-world challenging queries happening at night, while some others occur in different scales (from nearby views or from distant views). There is also a commonsense that some queries are blurred due to mobile capturing. We also selected some suboptimal queries that contain occlusions (objects or persons), as well as photos of partial landmark views. Figure 6 shows that, the compact descriptor from our vocabulary boosting can still well preserve the ranking precision, with comparisons to the Baselines (1)(2)(4). Figure 6 also investigates where our compact descriptor matches from the query photo to the reference database images, where the circles overlapped on the query and the first row of images denote the matched words.

mAP with Respect to Different Region: Figure 7 further shows the mAP variances in different regions, which shows that we can incrementally optimize the mAP in regions containing larger amounts of photos. It is also interesting that mAP is higher in the regions with a descriptor length of 100-200 bits, where indeed the majority of regions fall into this

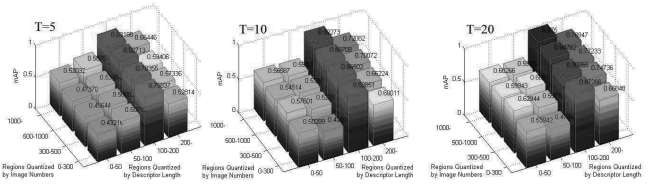

Figure 7: mAP variances in different regions, we draw two dimensional lattices to divided regions with respect to different image volumes and descriptor bits, then average the mAP for regions falling into each lattice.

category of 100-200 bits. Our extensive empirical study has demonstrated the more optimal performance for landmark search and returning additional information, with this setting of 100-200 bits.

What Compact Codewords are Learnt to Transmit: The learnt codebook is supposed to represent the most discriminative patches for a given query. Figure 8 further investigates which codewords are transmitted during the query. By visualizing the word centroid in Figure 8, we can see that different queries produce different codewords, where the most discriminant words are determined based on the resulting compact codebook to represent the query's visual content.

5 Conclusions and Future Works

In this paper, we propose to learn a compact visual descriptor by combining both visual content and geographical context, which has been deployed for mobile landmark search. We focus on its compactness for a low bit rate wireless transmission, as well as its computational efficiency for embedded feature extraction in mobile devices. We achieve both goals in an iterative optimization framework between geographical segmentation and descriptor learning. Our descriptor has been developed in both HTC DESIRE G7 and iPhone4 mobile phones, which outperforms state-of-the-art works in [Nister *et al.*, 2006][Chen *et al.*, 2009][Chandrasekhar *et al.*, 2009a][Jegou *et al.*, 2010], using one million Web landmark images covering Beijing, New York City, and Barcelona.

We envision the promising usage of context aware compact descriptors in the state-of-the-art research efforts as well as the emerging mobile search applications in industry. The context aware descriptors possibly start more practical scenarios to achieve real world extremely low bit rate transmission in WLAN or 3G s network. In addition, our proposed compact descriptor would be at the very beginning of a significant and potentially huge activity in research, development, and sstandardization of mobile search as well as mobile reality augmentation applications.

In addition, our efforts are closely targeted at the emerging MPEG Compact Descriptor for Visual Search (CDVS) standard. We think that this standardization and relevant research may stand as one of typical examples that integrate intelligent AI technique into embedded mobile platform. Nowadays, MPEG CDVS activity has attracted arising industry interests from quite a few smartphone or imaging chip companies like Nokia, Qualcomm, Aptina, NEC, *etc.*

Figure 6: Descriptor robustness against illumination changes, scale changes, blurred, occlusions, and partial queries. Each left photo is the query, each line of results corresponds to an approach. Top: Vocabulary Boosting; Middle: Original BoW or Tree Histogram Coding; Bottom: IDF Thresholding (top 20%). Based on the proposed compact descriptors, the spatial matching between each query (left photo) and the retrieved images (the top row) are illustrated by color circles (Different colors denote different codewords. Best view in color.).

6 Acknowledgements

This work was supported in part by the National Basic Research Program of China under contract No. 2009CB320902, in part by grants from the Chinese National Natural Science Foundation under contract No. 60902057 and 61071180, and in part by the CADAL Project Program.

References

[Chen *et al.*, 2009] D. Chen, S. Tsai, *et al.* Tree histogram coding for mobile image matching. *DCC*. 2009.

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

[Chandrasekhar *et al.*, 2009a] V. Chandrasekhar, G. Takacs, D. Chen, et. al. CHoG: Compressed histogram of gradients a low bit-rate feature descriptor. *CVPR*. 2009.

[Chandrasekhar *et al.*, 2009b] V. Chandrasekhar, G. Takacs, D. Chen, et. al. and B. Girod. Transform coding of image feature descriptors. *VCIP*. 2009.

[Nister *et al.*, 2006] D. Nister and H. Stewenius. Scalable recognition with a vocabulary tree. *CVPR*. 2006.

[Irschara *et al.*, 2009] A. Irschara, C. Zach, et al. From SFM point clouds to fast location recognition. *CVPR*. 2009.

[Schindler *et al.*, 2007] G. Schindler and M. Brown. City-scale location recognition. *CVPR*. 2007.

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

[Ke *et al.*, 2004] Y. Ke and R. Sukthankar. PCA-SIFT: A more distinctive rep. for local image descriptors. *CVPR*. 2004.

[Sivic *et al.*, 2003] J. Sivic and A. Zisserman et. al. Video Google: a text retrieval approach to object matching in videos. *ICCV*. 2003.

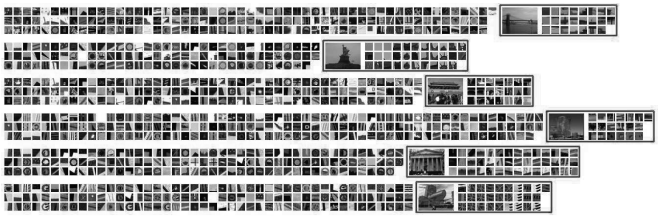

Figure 8: The learnt compact codebook and the extracted descriptors in exemplar queries in Barcelona. Left: the compact codebook in the query's assigned region; Middle: the query, where color highlights denote the detected descriptors on the query; Right: the transmitted words. We only transmit their occurrence index in practice. (Best view in color.)

[Philbin *et al.*, 2007] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Object retrieval with large vocabulary and fast spatial matching. *CVPR*. 2007.

[Jegou *et al.*, 2008] H. Jegou, M. Douze, and C. Schmid. Hamming embedding and weak geometric consistency for large scale image search. *ECCV*. 2008.

[Jurie *et al.*, 2005] F. Jurie and B. Triggs. Creating efficient codebooks for visual recognition. *ICCV*. 2005.

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

[Jiang *et al.*, 2007] Y.-G. Jiang, C.-W. Ngo, and J. Yang. Towards optimal bag-of-features for object categorization and semantic video retrieval. *CIVR*. 2007.

[Gemert *et al.*, 2009] J. Gemert, C. Veenman, A. Smeulders, and J. Geusebroek. Visual word ambiguity. *PAMI*. 2009.

[Kulis *et al.*, 2009] B. Kulis and K. Grauman. Kernelized Locality-Sensitive Hashing for Scalable Image Search. *ICCV*. 2009.

[Moosmann *et al.*, 2006] F. Moosmann, B. Triggs, and F. Jurie. Fast discriminative visual codebooks using randomized clustering forests. *NIPS*. 2006.

[Mairal *et al.*, 2008] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman. Supervised Dictionary Learning. *NIPS*. 2008.

[Lazebnik *et al.*, 2009] S. Lazebnik and M. Raginsky. Supervised learning of quantizer codebooks by information losss minimization. *PAMI*. 2009.

[Stauffer *et al.*, 2000] C. Stauffer and W. Grimson. Learning patterns of activity using real-time tracking. *PAMI*. 2000.

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

[Tsai *et al.*, 2010] S. Tsai, D. Chen, et. al. Location coding for mobile image retrieval. *MobileMedia*. 2010.

[Salton *et al.*, 1988] G. Salton *et al.* Term-weighting approaches in text retrieval. *Info. Proc. and Management*. 1988.

[Yuri *et al.*, 2010] Yuri Reznik. Compact Descriptors for Visual Search: Applications and Use Scenarios *Requirements Subgroup*. MPEG N11529. 2010.

[Ji *et al.*, 2011] R. Ji, L.-Y. Duan, J. Chen, H. Yao, W. Gao. A low bit rate vocabulary coding scheme for mobile landmark search. *ICASSP*, 2011.

[Chen *et al.*, 2011] J. Chen, L.-Y. Duan, R. Ji, H. Yao, W. Gao. Soring local descriptor for low bit rate mobile visual search. *ICASSP*, 2011.