Learning Compact Visual Descriptors for Low Bit Rate Mobile Landmark Search

Ling-Yu Duan, Jie Chen, Rongrong Ji, Tiejun Huang, Wen Gao

■ *Along with the ever-growing computational power of mobile devices, mobile visual search has undergone an evolution in techniques and applications. A significant trend is low bit rate visual search, where compact visual descriptors are extracted directly over a mobile and delivered as queries rather than raw images to reduce the query transmission latency. In this article, we introduce our work on low bit rate mobile landmark search, in which a compact yet discriminative landmark image descriptor is extracted by using a location context such as GPS, crowd-sourced hotspot WLAN, and cell tower locations. The compactness originates from the bag-of-words image representation, with offline learning from geotagged photos from online photo-sharing websites including Flickr and Panoramio. The learning process involves segmenting the landmark photo collection by discrete geographical regions using a Gaussian mixture model and then boosting a ranking-sensitive vocabulary within each region, with "entropy"-based feedback on the compactness of the descriptor to refine both phases iteratively. In online search, when entering a geographical region, the code book in a mobile device is downstream adapted to generate extremely compact descriptors with promising discriminative ability. We have deployed landmark search apps to both HTC and iPhone mobile phones, accessing a database of a million scale images in typical areas like Beijing, New York, and Barcelona, and others. Our descriptor outperforms alternative compact descriptors (Chen et al. 2009; Chen et al., 2010; Chandrasekhar et al. 2009a; Chandrasekhar et al. 2009b) by significant margins. Beyond landmark search, this article will summarize the MPEG standarization progress of compact descriptor for visual search (CDVS) (Yuri et al. 2010; Yuri et al. 2011) toward application interoperability.*

In recent years, mobile devices, such as smart camera phones and tablet PCs, have shown great potential for visual search, thanks to the integrated functionality of high-resolution embedded cameras, powerful CPUs, 3G/WI-FI wireless connections, color displays, and natural user interfaces. Emerging applications of mobile visual search and augmented reality include landmark search, product search, CD or book cover search, location recognition, and scene retrieval. As a popular application scenario, searching landmarks is one of the challenging tasks and has attracted great interest in both academic research and industrial practices. Existing mobile landmark search systems are deployed in the client-server architecture. The server end maintains a scalable, nearly duplicate visual search system, which is typically based on approximate visual matching techniques such as bag of words or hashing (Nistér and Stewénius 2006; Irschara et al. 2009; Schindler, Brown, and Szeliski 2007). To speed up similarity search, landmark photos in reference databases are in the form of an inverted index. In online search, a snapped landmark image is sent through a wireless link to the server, where a nearly duplicate visual search is conducted to identify the best-matched landmark in databases. Geographical location as well

Copyright © 2013, Association for the Advancement of Artificial Intelligence. All rights reserved. ISSN 0738-4602

Figure 1. The Developed Mobile Landmark Search System (Client App).

This system embeds the extraction of compact visual descriptors in the HTC Desire G7 mobile phone.

as tourist or other recommended information is subsequently returned to the mobile user.

The upstream transmission of query photos is subject to the bandwidth constraints of the wireless link. Undoubtedly, lengthy delivery latency may degenerate the user experience significantly. However, with the fast-increasing processing power in mobile devices, sending an entire image is unnecessary, as feature extraction and compression can be executed on mobile devices. To reduce the query transmission latency, the visual descriptor is supposed to be compact and discriminative. This technical trend has received dedicated efforts in MPEG standardization (that is, compact descriptor for visual search [CDVS] in MPEG [Yuri et al. 2010]). With evidence from research on low bit rate landmark visual search, our discussion will be extended to the ongoing MPEG CDVS standardization as well.

Indeed, extracting compact descriptors has a long history in the computer vision community; for instance the attempts to reduce the dimensions of local or global features (Ke and Sukthankar 2004; Jegou et al. 2010). However, as detailed later, in reference to low bit rate mobile landmark search, the previous local descriptors such as SIFT (Lowe 2004), SURF (Bay, Tuytelaars, and Van Gool 2006), or PCA-SIFT (Ke and Sukthankar 2004) cannot work well to meet the requirements in both descriptor compactness and extraction efficiency. Instead, recent attempts propose to extract much more compact visual descriptors (Chen et al. 2009; Chen et al. 2010; Chandrasekhar et al. 2009a; Chandrasekhar et al. 2009b), say tens of bits per local descriptor, to reduce the query delivery latency. However, existing works are solely based upon visual contents to compress descriptors, regardless of rich mobile context such as GPS, crowd-sourced WLAN hotspots, or cell tower locations. In this work, we achieve descriptor compactness through contextual learning, with additional concerns regarding the mobile end extraction complexity.

Coming up with a solution of this contextual learning for mobile landmark search, we explore the combination of mobile context and visual statistics in each geographical region to learn a com-

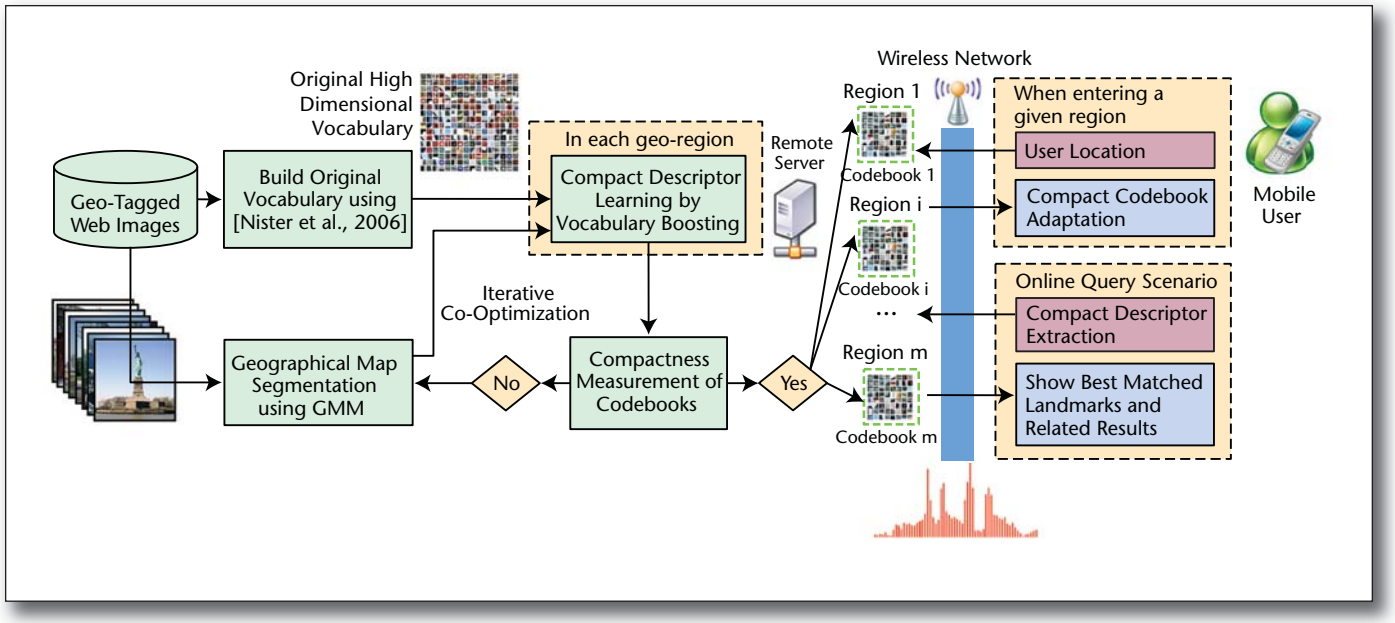

Figure 2. The Mobile Visual Landmark Search System

The proposed contextual-learning-based compact visual landmark descriptor extraction framework toward low bit rate mobile landmark search, which embeds descriptor extraction into the mobile end.

pact descriptor from geotagged reference photo collections.[1] First, we propose a geographical segmentation of geotagged photos based on the Gaussian mixture model (Stauffer and Grimson 2000); second, we introduce a vocabulary-boosting scheme to learn a compact descriptor in each region, which simulates a set of landmark queries from this region and learns a compact code book to maintain the search precision from an original visual vocabulary (Sivic and Zisserman 2003; Nistér and Stewénius 2006). With this code book, a compact bag-of-words-based descriptor is generated for a given query. However, due to imprecise segmentation, learning compact descriptors separately in each individual region cannot guarantee a global optimum. To address this issue, we propose to iterate the twin stages of content-aware geographical segmentation and the vocabulary boosting to reinforce each other. Figure 2 shows the mobile visual landmark search system with contextual-learning-based compact visual descriptors embedded in the mobile end. Although geoinformation is used, the context-based partition as well as the partition-based compact code books is generic in a sense.

In practice, once a mobile user enters a region, the server transmits a downstream supervision (that is, a compact code-word boosting vector) to teach the mobile device by linearly projecting the original high-dimensional vocabulary into a compact code book through this boosting vector. Given a query, instead of a high-dimensional code-word histogram, an extremely compact histogram is redirected to transmit.

Related Work and Challenges

Undoubtedly, geolocation applications such as mobile landmark and location search have received a wide range of attention from both academe and industry. For instance, mobile location recognition (Irschara et al. 2009; Philbin et al. 2007; Crandall et al. 2009; Zhang and Kosecka 2006; Shao et al. 2003), mobile landmark identification, online photograph recommendation (Hays and Efros 2003; Li et al. 2008; Zheng et al. 2009), and content-based advertising (Liu et al. 2009). In particular, Google Project Glass attempts to develop an augmented reality head-mounted display (resembling a pair of eyeglasses), allowing the integration of phones, GPS, and cameras to display augmented information on the screen.

Most existing geolocation search systems follow a client-server architecture. Take landmark search as an example. The remote server maintains a landmark database. In online search, mobile users take a query photo, which is transmitted to remote servers to identify its corresponding landmark through visual matching. Based on a reference database, the server returns search results including the mobile user's geographical location, photograph viewpoints, recommendations for tourism, or other value-added information.

Generally speaking, in most existing mobile visual search systems (such as Google Goggles, Nokia Point and Find, and others), query photos are delivered over a bandwidth-constrained wireless link. User experience heavily depends on how

much data is transmitted. It is easy to imagine that sending the entire photo is time consuming and is not necessary indeed. The ever-growing computational power motivates the research efforts to extract visual descriptors directly on a mobile device (Chen et al. 2009; Chandrasekhar et al. 2009a; Makar et al. 2009). To the best of our knowledge, sending a compressed bag of features requires 2–4 KB descriptors per query in the state of the art (Chen et al. 2009; Chen et al. 2010).

Compared to previous work on compact local descriptors, for example, SURF (Bay, Tuytelaars, and Van Gool 2006), GLOH (Mikolajczyk and Schmid 2005), PCA-SIFT (Ke and Sukthankar 2004), and MSR descriptors (Hua, Brown, and Winder 2007), recent works by Chen et al. (2009), Chandrasekhar et al. (2009a), and Makar et al. (2009) have attempted to address much lower bit rate visual descriptors, especially targeting mobile visual search.

A review on compact descriptors can be found in Girod et al. (2011). The first typical category works on the quantization of local descriptors to reduce the size of each local descriptor. For instance, Chandrasekhar et al. (2009a) proposed a compressed histogram of gradient (CHoG) for compact visual descriptors, which adopts both Huffman tree and Gagie tree to compress each local feature into approximately 60 bits. Chandrasekhar et al. (2009b) also proposed to compress the SIFT descriptor with Karhunen-Loeve transform, yielding approximate 2 bits per SIFT dimension. Tsai et al. (2009) further proposed to code the spatial layout of interest points for the subsequent image matching to improve search precision by reranking. Rather than sending a query photo, sending compact local descriptors (Chandrasekhar et al. 2009a; Chandrasekhar et al. 2009b; Tsai et al. 2009) is much more effective in reducing bit budget and transmission latency. For instance, with a normal local feature detector setting (Mikolajczyk and Schmid 2005), about 1000 interest points will be detected per image; the overall amount of feature data is about 8 KB, much less than the size of the query image (typically more than 100 KB with JPEG compression).

Different from directly compressing local descriptors, the second category is to further compress the (vector) quantized bag-of-words signature (Chen et al. 2009; Chen et al. 2010; Ji et al. 2011). For instance, Chen et al. (2009) proposed a tree histogram coding (THC) scheme to compress the sparse bag-of-words signature, which encodes the position difference of nonzero bins for high compression rates. THC yielded an approximate 2 kilobytes of code per image for a vocabulary with one million words, much less than directly sending CHoG descriptors (Chandrasekhar et al. 2009a) (more than 6 KB). To maintain a scalable visual code-book-based retrieval system, Chen et al. (2010) further compressed the inverted indices of the vocabulary tree model (Nistér and Stewénius 2006) with arithmetic coding for memory reduction at the server. Recently, Ji et al. (2011a) proposed learning a compact descriptor adaptively within different geographical regions for mobile landmark search, which enables a region-specific landmark descriptor.

The aforementioned two categories may be unified from the production-quantization (Gray and Neuhoff 1998) point of view. In product quantization, an input vector is divided into k segments, and those k segments are independently quantized using k subquantizers. Each compressed descriptor is thus represented by k indexes comprising the nearest code words of k segments. When $k = 1$, the first category degenerates to the second category. Usually, quantization tables in the first category are smaller.

Beyond diverse quantization methods (Gray and Neuhoff, 1998), maintaining sufficient discriminability as well as desirable descriptor compactness is essentially a sort of trade-off; that is, an elegant descriptor design comes from the optimization of both factors. In the following paragraphs, we summarize key challenges in the state of the art, which thereby motivates the proposed contextual learning of landmark descriptors.

Location Insensitive Compression

Existing compact descriptors rely on visual statistics solely. But the rich and cheaply available location cues are left unexploited. Such location contexts have been widely available from either mobile devices or landmark photo collections. That is, the landmark descriptor should be location sensitive, taking into account where the query happens.

Unscalable Landmark Description

Existing compact descriptors are less scalable in length with respect to regions, each of which would be supposed to maintain its own discriminability. The scalability may depend on the visual complexity of landmark images at a given location. For instance, the descriptors in a location containing multiple landmarks with diverse appearances could be less compact.

One-Way Coding Transmission Mode

Existing compact descriptors, directly extracting features for upstream query delivery, do not employ the two-way communication mode of mobile devices. However, given a batch of queries, there is never a constraint that only upstream query transmission is allowed before a server performs retrieval and returns downstream results. For instance, by leveraging the location of a mobile

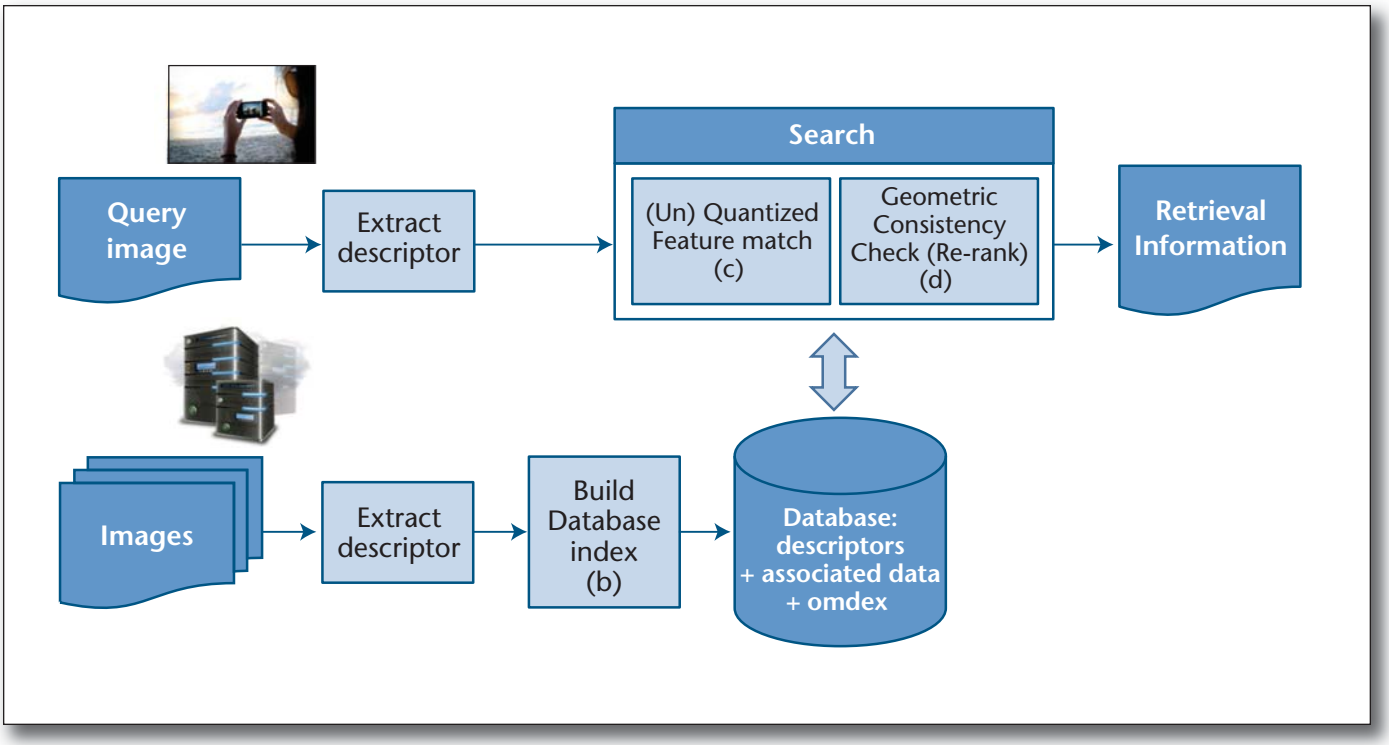

Figure 3. Flowchart of Visual Search and the Basic Components of Feature Extraction, Indexing, Matching, and Reranking.

Visual search applications work in a client-server architecture, that is, a large reference image database is set up in a remote server and a client sends queries to the server for visual search. Feature extraction (a) and indexing (b) over database images are performed at an offline stage. At online stage, the client end extracts features and sends them to the server. Query features are subsequently matched against database features in the server end (c). The retrieval results are then verified with a geometric reranking (d). Finally, either top-ranked images or their relevant information is delivered to the client end.

user, the remote server may predeliver a compact downstream supervision facility to "teach" the mobile how to extract compact and discriminative descriptors.

To address these challenges, we attempt to learn the compact descriptor based on the mobile context.[2] In the landmark search scenario, we emphasize exploiting the location context, coming up with a novel extremely compact descriptor that is location sensitive, scalable, and of two-way coding mode.

Visual Search Preliminary

Searching for a specific object in a large collection of images has been a long-standing research problem, with a wide range of applications in location recognition and product search. Much research has been done on different modules such as image representation, matching, indexing, retrieval, and geometric reranking. Figure 3 gives a baseline flowchart for visual search. Local features are extracted from database images at servers. The server searches for relevant images based on the visual similarity between a query and database images, and inverted indexing is often employed to speed up the search process. Local features are proven to be particularly important in visual search.

Using a brute force method to match query features with database features is infeasible for large-scale databases, and a feature index (for example, inverted indexing) based on visual vocabulary is needed to improve search efficiency (see figure 3b). Other alternative solutions include approximate nearest-neighbor search, such as *k*-D tree or locality-sensitive hashing. In addition, feature-space quantization schemes, such as *k*-means and scalable vocabulary tree, have been widely used for scalable image search, in which two features are considered the same word when they fall into the same cluster.

In online search, local features in a query image are extracted and used to search for the local features' nearest neighbors based on the feature database from reference images. Database images containing the nearest neighbors are efficiently collected based on index files and are accordingly ranked by the similarity score (figure 3c). Finally, the top returned images are reranked through geometric consistency checking in which the location of local descriptors is considered as well (figure 3d).

In particular, as a typical approach to scalable

indexing, learning visual vocabulary usually resorts to unsupervised vector quantization such as *K*-means (Sivic and Zisserman 2003), which partitions local feature space into code-word regions. An image is then represented by a bag-of-words histogram, where each bin counts local features being quantized into the corresponding code word. Many vector-quantization-based vocabularies, such as vocabulary tree (Nistér and Stewénius 2006), approximate *k*-means (Philbin et al. 2007), Hamming embedding (Jegou, Douze, and Schmid 2008; Jurie and Triggs 2005; Jiang, Ngo, and Yang 2007; Philbin et al. 2007; Jegou et al. 2010; Ji et al. 2009; Ji et al. 2010) has been reported. Hashing-based approaches, such as locality sensitive hashing and its kernelized version, are another alternative (Kulis and Grauman 2009). In addition, the work of Jiang, Ngo, and Yang (2007), Jegou, Douze, and Schmid (2008), Philbin et al. (2007), and Gemert et al. (2009) attempts to deal with code-word uncertainty and ambiguity, for example, Hamming embedding (Jegou, Douze, and Schmid 2008), soft assignments (Philbin et al. 2007), and kernelized code book (Gemert et al. 2009). Recent work has employed semantics or category labels to supervise the vocabulary construction (Moosmann, Triggs, and Jurie 2006; Mairal et al. 2008; Lazebnik and Raginsky 2009).

The Working Pipeline

The system (figure 2) developed in this article works in a bidirectional manner in terms of information exchange. To search the landmark location, a raw bag-of-words signature is extracted in the mobile end while the context tags such as GPS or cell tower locations are obtained as well. Subsequently, the following operations are performed step by step.

Phase 1 is a region selection operation in the mobile end. Its input can be side information of mobiles like GPS or cell tower tags that are available at the mobile end directly and that are used to locate the current query, namely, the geographical region in a given city.

Phase 2 extracts local features, quantizes them into visual words, and forms a bag-of-words-based "topical" descriptor, which is binarized into an occurrence (hit/nonhit) histogram followed by Huffman coding.

Phase 3 transmits the encoded signature together with side information to a remote server.

Phase 4 decodes the topical descriptor at the server to recover the original bag of words, which is then combined with the region-specific compact vocabulary **f** to search for duplicate or near-duplicate images.

Toward compact visual descriptors, this working pipeline essentially employs a contextual learning to optimize the vocabularies (in terms of scale and retrieval performance of test queries) with respect to different locations. In the following subsection, we will clarify the input and output of the contextual learning, as well as the setup of contextual learning goals.

Input and Output of the Contextual Learning

The inputs are twofold, first, the original high-dimensional visual vocabulary **V** trained from a very large image collection, and second, the contextual (side) information (namely GPS in this article) *G* for each geotagged image.

The outputs are twofold as well. First, segmenting the geographical map of geotagged reference images into discrete regions; and second, learning a compact vocabulary within each geographical region to generate online extremely compact landmark descriptors for low bit rate visual search.

Learning Goal

Given database images $\mathbf{I} = \{I_i\}_{i=1}^{n}$, we extract offline *n* bag-of-words histograms (Nistér and Stewénius 2006; Sivic and Zisserman 2003) $\mathbf{V} = \{V_i\}_{i=1}^{n}$, which are high-dimensional, say 0.1–1 million in state-of-the-art settlings (Nistér and Stewénius 2006).[3] Note that all database images are tagged with GPS coordinates as $\mathbf{G} = \{Lat_i, Long_i\}_{i=1}^{n}$.

We aim first to learn a geographical segmentation $\mathbf{S} = \{S_j\}_{i=1}^{m}$ to partition $\mathbf{I} = \{I_i\}_{i=1}^{n}$ into *m* regions, which attempts to employ the local context to achieve descriptor compactness to an extreme; and second, learn a code book $\mathbf{U}_j \in \mathbb{R}_k$ for compact descriptor extraction in each S_j from $\mathbf{V} \in \mathbb{R}_n$ such that $k \ll n$, which is updated online in the mobile device once the mobile user enters S_j.

Indeed, the above coupled goals are a chicken and egg problem: On one hand, we expect the code book $\mathbf{U}_j$ is as compact as possible in each S_j. On the other hand, the compactness depends on how properly the image collection in **I** is segmented into S_j. However, such segmentation is often imprecise, especially in 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. In other words, the optimization of descriptor compactness is local in each region, rather than global among all the regions of the entire image database. Ideally, we aim jointly to learn both the optimal region segmentation and the compact description to minimize:

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

where $|U_i|$ denotes the descriptor length of the *i*th sampled query image (in total *n'*) falling into region S_j; the constraint denotes the retrieval pre-

cision loss ($Loss(P_{S_j})$) in each region, which will be revisited later in this article. However, we cannot perform 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 over the entire image data set of a city.

Compact Descriptor Learning

In this section, we introduce the learning of compact descriptors including the segmentation of the geographical map and vocabulary boosting, as well as the iterative optimization of two separate stages.

Geographical Map Segmentation

To reduce the impact of visual diversity on descriptor compactness, we first perform geographical map segmentation. The assumption is that the retrieval task in smaller regions may alleviate the requirements of larger descriptors. We employ the Gaussian mixture model (GMM) (Stauffer and Grimson 2000) to segment **I** into **S.** We assume the geographical distribution of photos is drawn from m landmark regions, and denote the ith component as w_i, with mean vector μ_i. We regard those photos belonging to the ith component as generated from the ith Gaussian with mean μ_i and covariance matrix Σ_i, following a normalized distribution $N(\mu_i, \Sigma_i)$, with the probability of each ith component P_i.

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

$$p(y = i \mid x) = \frac{p(x \mid y = i)P(y = i)}{p(x)} \tag{2}$$

where $p(x|y = i)$ is the probability of photo x based on the condition that photo x comes from the ith component, following a normalized distribution:

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

We adopt an expectation maximization (EM) to perform segmentation. We will revisit P_i in the iterative co-optimization section to learn geographical segmentation and compact descriptors in a joint optimization manner.

Descriptor Learning Through Vocabulary Boosting

Toward efficient visual search in a million scale database, the scalable vocabulary tree (SVT) (Nistér and Stewénius 2006) is well exploited in previous works (Chen et al. 2009; Chen et al. 2010; Irschara et al. 2009; Schindler, Brown, and Szeliski 2007). SVT uses hierarchical K-means to partition local descriptors into quantized code words. An H-depth SVT with B-branch produces $M = B^H$ code words, and a typical setting is $H = 5$ and $B = 10$ (Nistér and Stewénius 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 the vocabulary hierarchy to find out the nearest code word, which converts $\mathbf{L}(q)$ to a bag-of-words histogram $\mathbf{V}(q) = [V_1(q), ..., V_M(q)]$.

Given a query image, visual search may be formulated as the problem of minimizing a loss function with respect to the ranking position $R(x)$) of retrieved images I_x (bag-of-words feature $\mathbf{V}(x)$) as follows:

$$Loss_{Rank} = \sum_{x=1}^{n} R(x)\mathbf{W}_x \| \mathbf{V}(q), \mathbf{V}(x) \|_{Cosine} \tag{4}$$

where tf-idf (term frequency–inverse document frequency, that is, IF-IDF) weight $\mathbf{W}_x$, together with the cosine distance,[4] is applied to measure the relevance between query and retrieval images in a similar way as the document retrieval method (Salton and Buckley 1988). Intuitively, by minimizing $Loss_{Rank}$, more relevant images would be assigned to a higher rank. For example, when the cosine distance is zero, the returned image is irrelevant, which means that $R(x)$ may be assigned a bigger value and vice versa.

By dealing with each image as a document, the TF-IDF weight is computed as follows:

$$\mathbf{W}_x = [\frac{n_1^x}{n^x} \times log(\frac{n}{n_{V_1}}), ..., \frac{n_M^x}{n^x} \times log(\frac{n}{n_{V_M}})] \tag{5}$$

where n^x denotes the number of local descriptors in I_x; $n_{V_i}(x)$ the number of local descriptors in I_x being quantized into V_i; n_i^x / n^x, the term frequency of V_i in I_x, and $log(n / n_{V_i})$, the inverted document frequency of V_i.

Simulating User Queries to Learn Vocabulary

Given a region containing n' landmark photos $[I_1, I_2, ..., I_{n'}]$, we sample a subset of geotagged photos $[I'_1, I'_2, ..., I'_{n_{sample'}}]$ to simulate user queries. Through reducing the loss function from these pseudoqueries, we attempt to optimize the compact set of code words. Currently, query photos are randomly sampled from the region; however, the sampling strategy may be customized in different scenarios (say using a user query log). For clear explanation, we denote the collection of retrieval ranking lists of simulation user queries as follows:

$$\begin{aligned} Query(I'_1) &= [A_1^1, ..., A_R^1] \\ &\cdots \\ Query(I'_{n_{sample}}) &= [A_1^{n_{sample}}, ..., A_R^{n_{sample}}] \end{aligned} \tag{6}$$

where A^j_i is the ith returning of the jth query.

Ideally, a compact descriptor is supposed to

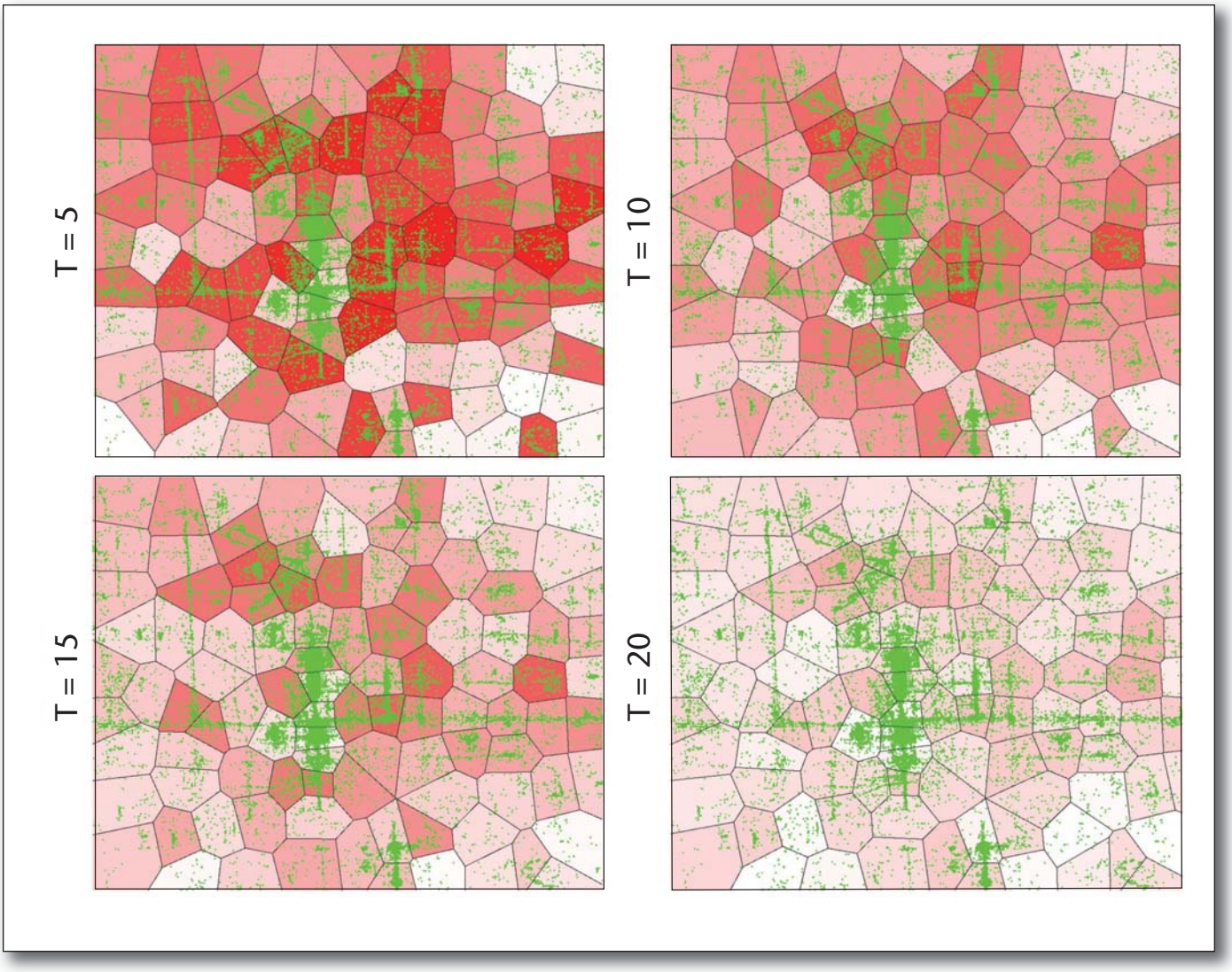

Figure 4. The Geographical Visualization of the Descriptor Compactness in Beijing City.

Shown through iterative co-optimization (T = 1 to 20). We normalize the descriptor length by a min versus max ratio, where the ratio is represented by the saturation of the color red. Green dots indicate the distribution of geotagged photos. Less saturation on the map means more optimal descriptors.

maintain the original ranking list $[A^j_1, A^j_2, ..., A^j_R]$ as much as possible for the jth query based on the (uncompressed) raw descriptors. In the subsequent learning process, the joint set of pseudoqueries and their retrieval results are considered as the ground truth to minimize the total ranking loss through training. In other words, the learner is to generalize from the training query instances in order to produce compact and effective descriptors in new queries.

Location-Aware Vocabulary Boosting

Compact descriptor learning may be formulated as an AdaBoost-based code word selection. Each single code word acts as a weak learner, and the learning target is to minimize the ranking discriminability loss with a minimum coding length.

We first define $[w_1, ..., w_{nsample}]$ as an error weighting vector to n_{sample} query images in region S_j, which estimates the loss of ranking consistency in the current word selection. We then define the encoded vocabulary as $\mathbf{U}_j \in \mathbb{R}_k$ for region S_j, which is obtained from raw bag-of-words feature $V \in \mathbb{R}_m$ through $\mathbf{U}_j = \mathbf{M}^T\mathbf{V}$, where $M_{M \times K}$ is a bag-of-words feature dimension reduction transform from $\mathbb{R}_m$ to $\mathbb{R}_k$.

Concretely speaking, we resort to a learning process to spot effective code words incrementally. In a greedy strategy, at each stage the boosting iteratively selects the top influential words in terms of the retrieval ranking loss over pseudoqueries. This strategy does not need to find a best solution but terminates in a reasonable number of steps. In a

formal way, the bag-of-words dimension reduction transform may be represented by a diagonal matrix $\mathbf{MM}^T$, in which nonzero diagonal positions denote the selection of code words.

Assuming that at the tth iteration, we have determined $(t-1)$ nonzero diagonal elements in MMT, corresponding to the selection of $(t-1)$ code words. At the next, to determine the tth discriminative code word, we estimate the ranking loss at the current code words setting of $\mathbf{MM}^T$:

$$Loss(I'_i) = w_i^{t-1} \sum_{r=1}^{R} R(A_r^i)\, \mathbf{W}_{A_r^i} \| \mathbf{M}^{t-1}\mathbf{V}(I'_i), \mathbf{V}(A_r^i) \|_{Cosine} \quad (7)$$

where $i \in [1, n_{sample}]$; $R(A_r^i)$ denotes the renewed rank of the original rth retrieved image for query I'_i when performing retrieval with selected code words; w_i^{t-1} denotes the error weight of sample i at the $(t-1)$th stage. By summing up the ranking loss of n_{sample} queries, we have the rank loss:

$$Loss_{Rank} = \sum_{i=1}^{n_{sample}} Loss(I'_i) \quad (8)$$

At the $(t-1)$th stage, one new code word U_t is selected by minimizing the summed up ranking loss:

$$U_t = \arg\min_j \; Loss_{Rank} \quad (9)$$

The boosting process terminates when

$$\sum_{i=1}^{n_{sample}} w_i^t \leq \tau$$

Iterative Co-Optimization

In this subsection, we further investigate the problem of optimizing compact vocabularies across the set of regions. It is well known that mobile Internet systems have an inherent asymmetric communication channel between "servers" (powerful cloud search engine) and clients (weak mobiles), with faster communication from server to client than from client to server. Generally speaking, broadband connections have a download and upload bandwidth ratio typically between a factor of 5 and 15. To accomplish effective mobile visual search, the client and the server must exchange enough information. In particular, besides the objective of reducing the number of bits sent by the client, we also need to minimize the number of rounds of communication.

In the context of low bit rate mobile landmark search, we attempt to figure out the trade-off between the downstream vocabulary adaptation and the upstream descriptor delivery. It is easy to imagine that, in subdividing geographical regions, finer partition (smaller regions) would probably yield more compact code books whereas more frequent downstream adaptation could occur. Ideally, we aim to yield more compact descriptors; however, we have to consider how to merge nearby regions to reduce the rounds of downstream adaptation at a slight loss of descriptor compactness.

Hence, we propose the co-optimization process, involving the iterative phases of geographical segmentation as well as compact descriptor learning. Through the iterative co-optimization, we aim to minimize three important aspects: the number of bits sent by the mobile (compact descriptors), the number of bits sent by the server (vocabulary), and the number of rounds of communication (region-wise adaptation for a mobile user).

We introduce an uncertainty measurement (like entropy) to determine the necessity of longer or shorter descriptor length of a region, which is feedback to the segmentation stage to adjust the values of a priori probability P_i in the Gaussian mixture model, thereby refining the geographical segmentation at the Tth iteration. By using the size of vocabularies, the probability P_i is iteratively updated as:

$$P_i = -log\left(|\mathbf{U}_i| / |\mathbf{U}_{max}|\right) \quad (10)$$

A larger vocabulary size $|\mathbf{U}_i|$ leads to a lower a priori probability value, indicating that this "complex" region (a Gaussian mode) tends to be split; a smaller vocabulary size $|\mathbf{U}_i|$ leads to a higher a priori probability value, which means that its surrounding regions tend to be merged into this "simple" region to form a bit more complex region.

Figure 4 visualizes the process of city-scale iterative optimization of descriptor length involving the results of distinct phases of segmentation and learning (T = 1 to 20). The illustration of colors in regions intuitively tells that the descriptors' lengths in different geographical regions are being gradually minimized overall. In real-world scenarios, when a mobile user travels in a city and visits multiple landmarks, or multiple travelers visit different landmark regions, the overall bandwidth cost from downstream descriptor adaption and upstream query transmission would probably be minimized.

Quantitative Results

We collected more than 1 million geographical tagged photos from both Flickr and Panoramio photo-sharing webs, which cover typical areas including Beijing, New York City, and Barcelona. Based on the geographical map of each city, we chose 30 dense regions and 30 random regions.[5] As manually identifying all (near)-duplicated photos of a landmark is intensive, we invite volunteers to identify one or more dominant views of each landmark. All near-duplicated landmark photos to a given view are labeled according to its belonging and nearby regions. We sample 5 images from each region as queries, which finally forms in total 300 queries in each city.

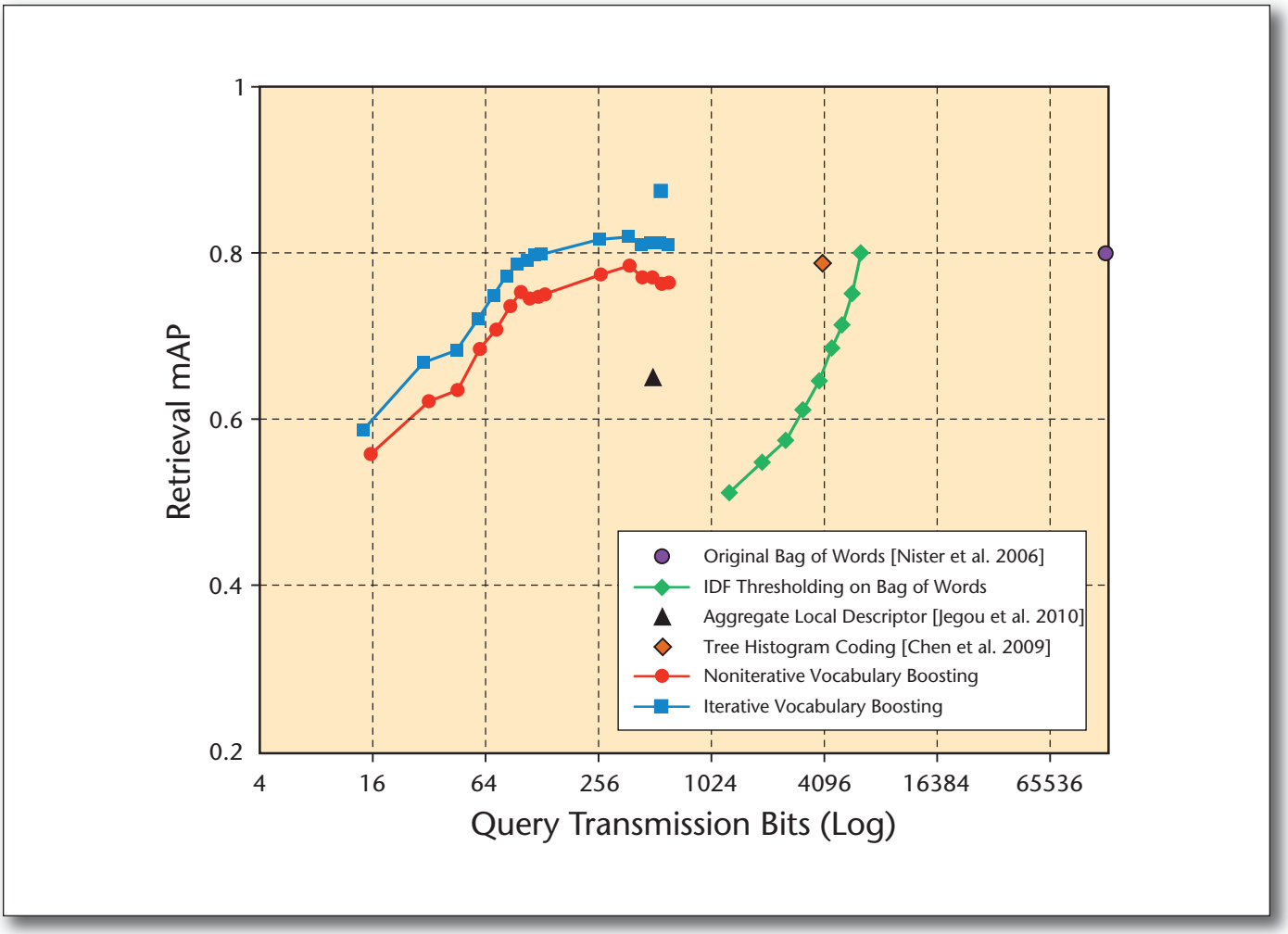

Figure 5. Compression Rate Versus Ranking Distortion of Our Descriptor Learning Comparing with the State of the Art.

Parameters and Evaluations

We employ SIFT (Lowe 2004) as local descriptors. A scalable vocabulary tree (Nistér and Stewénius 2006) is built to form the initial vocabulary $\mathbf{V}$, which generates a bag-of-words signature $\mathbf{V}(i)$ for each database image $\mathbf{I}_i$. We apply the identical vocabulary generated in Beijing to other cities. For each region, the boosting is conducted to learn $\mathbf{M}$ offline. We denote the hierarchical level as H and the branching factor as B. We have $H = 5$ and $B = 10$, producing approximately 0.1 million words. The mean average precision (*mAP*) is used to evaluate retrieval performance, which reveals its position-sensitive ranking precision in the top 10 positions.

Baselines

(1) *Original bag of words:* Transmitting the entire bag-of-words histogram has the lowest compression rate. However, it provides the performance upper bound in terms of mean average precision. (2) *IDF thresholding:* As a straightforward scheme, we transmit the IDs of code words with the highest IDF values (figure 5 tests 20 percent to 100 percent of code words) as an alternative vocabulary compression. (3) *Aggregating local descriptors* (Jegou et al. 2010): Jegou et al. (2010) adopted aggregated quantization to obtain a compact signature. Its output is a global compact representation by learning a model of local descriptors. (4) *Tree histogram coding*: Chen et al. (2009) used residual coding to compress the bag-of-words histogram, which is the closely related work. (5) *Without co-optimization:* To quantize our iterative co-optimization, we degenerate our approach by dismissing iterations between geographical segmentation and descriptor learning.

Rate Distortion Analysis

We compare the rate distortion with state-of-the-art works (Nistér and Stewénius 2006; Chen et al. 2009; Chandrasekhar et al. 2009a; Jegou et al. 2010) in figure 5. We achieve the highest compres-

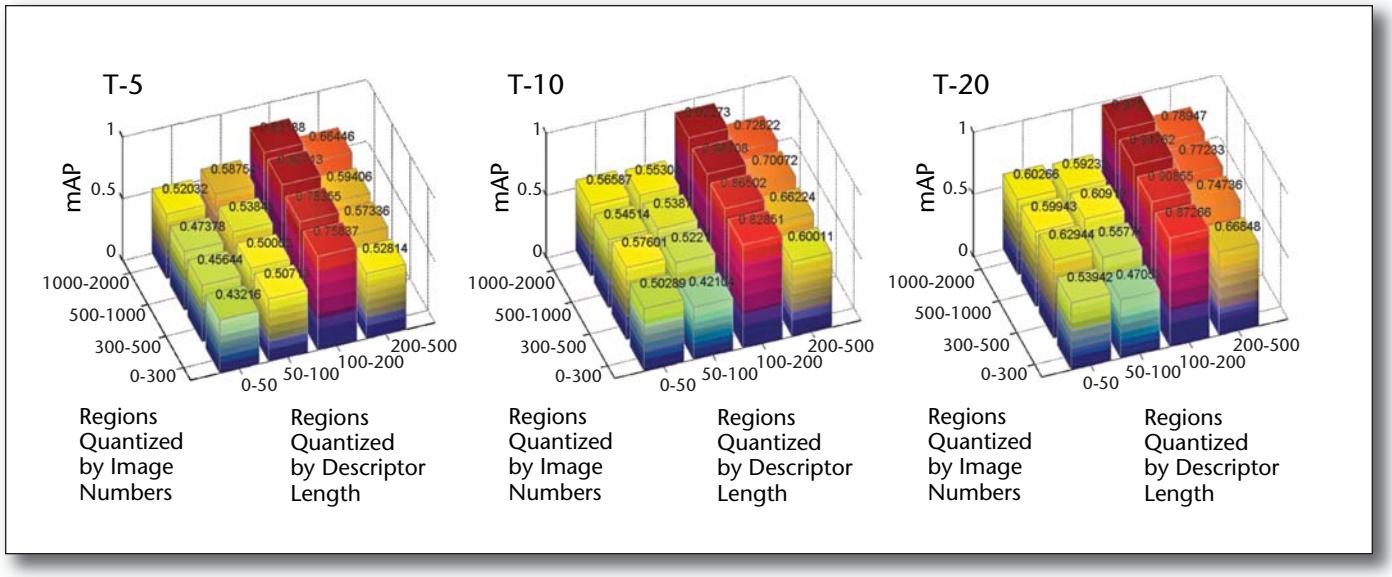

Figure 6. The Mean Average Precision Variances of Different Types of Regions.

Two-dimensional lattices are formed by the region statistics with respect to different image volumes and descriptor lengths in bits. The mean average precision is obtained by averaging the search results of all the test queries of those regions falling into each lattice.

sion rate with a given performance distortion (horizontal view), while maintaining the best mean average precision at a given compression rate (vertical view).

The Mean Average Precision with Respect to Different Regions

Figure 6 further indicates the mean average precision variances in different regions, in which the mean average precision in those regions containing many photos can be incrementally optimized. We can see that the mean average precision is higher in the regions with a descriptor length of 100–200 bits, where indeed the majority of regions fall into this category of 100–200 bits. Our extensive study has demonstrated that the optimal performance for landmark search is normally with the descriptor setting of 100–200 bits.

Energy Consumption Analysis

Energy conserving is critical for mobile applications. One interesting study is to compare the average mobile energy consumption in (1) extracting and sending compact descriptors, and (2) sending the original query image. In the 3G environment, we empirically test the maximum number of queries that the mobile can send before the battery is flat. A typical phone battery (HTC Desire G7) has a voltage of 4.0 V and a capacity of 1400 mAh (or 20.2 kilojoules). Hence, for 3G connections, the maximum capability of sending images (VGA size) is 20.2 kilojoules/52.4 joules = 385 total queries. However, for extracting and sending compact descriptors, up to 20.2 kilojoules/8.1 joules = 2494 queries are allowed.

Insights into Compact Descriptors

Descriptor Robustness and Matching Locations: We collect quite a few real-world challenging queries at night, as well as queries at different scales (close or distant views). We also select worse queries involving occlusions (objects or persons), and partial landmark views. Figure 7 shows that the compact descriptor from our vocabulary boosting can better preserve the ranking precision, compared to the baselines (1)(2)(4). Figure 7 illustrates how our compact descriptor matches the query photo to the reference database images, where circles indicate the code words of each query and the first row of images displays the matched words.

Descriptor Sensitiveness to Landmark Scale: To study the effects of landmark scale on the compact descriptor, we empirically categorize landmark scale by measuring the geographical distribution of reference images. Based on the maximum distance of the minimum containing region, we come up with three scales, small, medium, and large, as illustrated in figure 8. The typical distance for small scale is 0–12 meters, the medium scale is 12–30 meters, and the large scale is more than 30 meters. It is found that the performance of searching large-scale landmarks is much better than the medium and small scales, due to less background clutter and more distinguishing feature points. However, in some cases, the location-context-assisted search performance of small scale is better than large scale, as the GPS signal may be distorted around a large-scale landmark, while location context plays a relatively important role; therefore, small-scale landmarks may yield better results by context-assisted visual search.

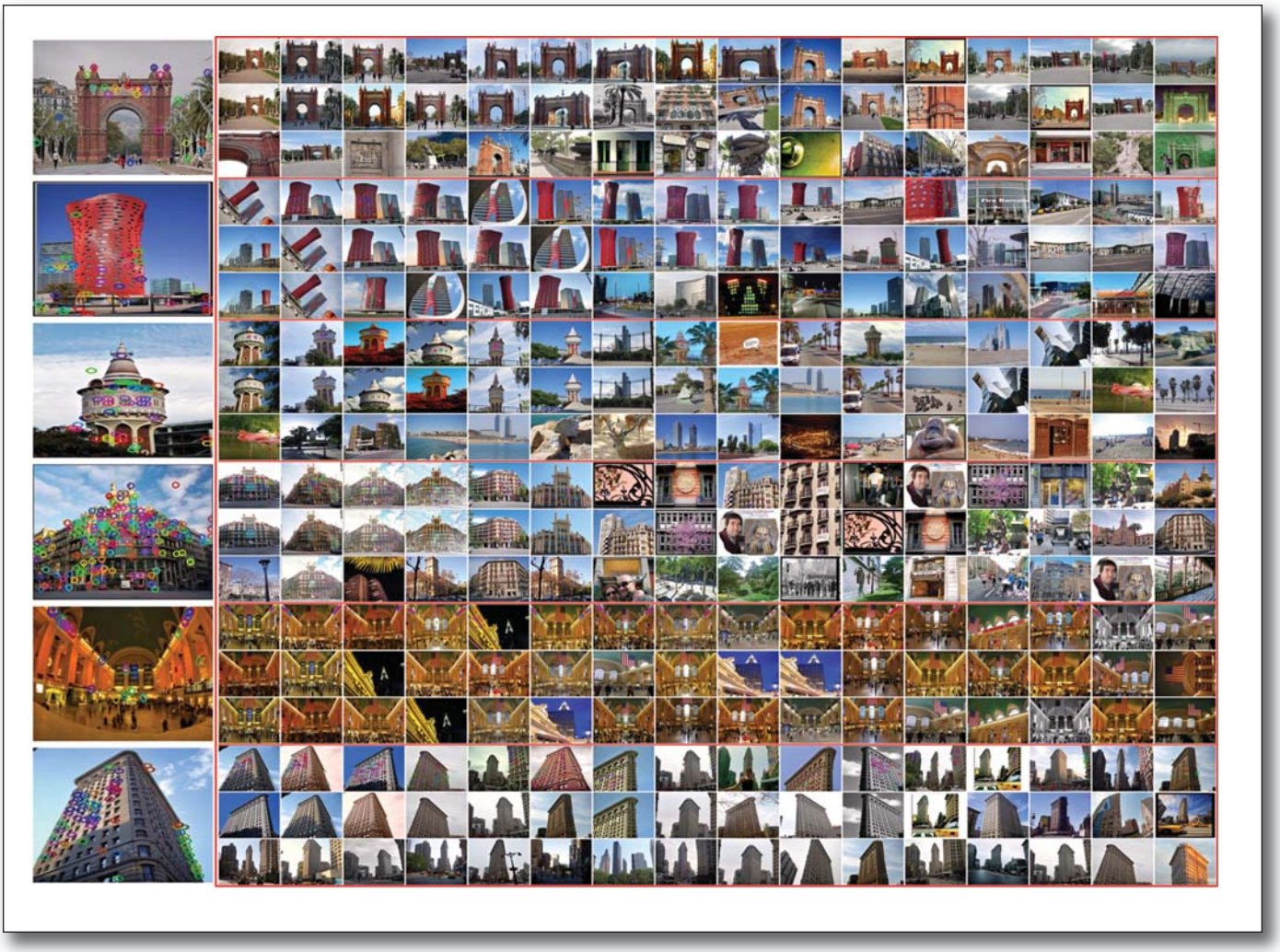

Figure 7. Exemplar Queries to Demonstrate the Descriptor Robustness Against Illumination Changes, Scale Changes, Blurriness, Occlusions, and Partial Queries.

In each photo the query is on the left, and each line corresponds to the retrieval results of an approach. *Top:* Vocabulary boosting; *Middle:* Original bag-of-words or tree histogram coding. *Bottom:* IDF thresholding (top 20 percent code words). The code-word matching between each query (left photo) and the retrieved images (the top row) are illustrated by circles (different colors denote different code words).

More quantitative study on the effects of landmark scale, shot size, and other photograph factors is beyond the scope of this article. Readers are referred to Peking University Landmarks benchmark (PKUBench) (Ji et al. 2011c), which involves 198 landmarks in the Peking University campus with sufficient coverage of mobile photographing variances.

What Learned Compact Code Words Are Transmitted

The learned code book is supposed to represent the most discriminative patches of the photos within each region. Figure 9 indicates which code words are transmitted actually to represent queries. Referring to the visualized word *centroid* in figure 9, different queries generate different code words, where the most discriminative words are selected based on the compact code book determined for each region.

As compact code books rely on the boosting process of pseudoqueries over a reference database, the selected code words for new landmarks (that the system has not encountered) may be (much) less optimal. In practice, GPS context can alleviate such negative effects. How to properly identify and represent new landmarks based on boosting-based code books will be included in our future work.

Beyond Landmark Search

In addition to a landmark data set, we have conducted extensive experiments on various data sets (PKUBench, UKBench [Nistér and Stewénius 2006]), to demonstrate that the context-assisted MCVD descriptor described by Ji et al. (2011d),

Figure 8. Exemplar Images of Different Scales in PKUBench, from Top to Bottom.

(a) Small; (b) Medium; (c) Large.

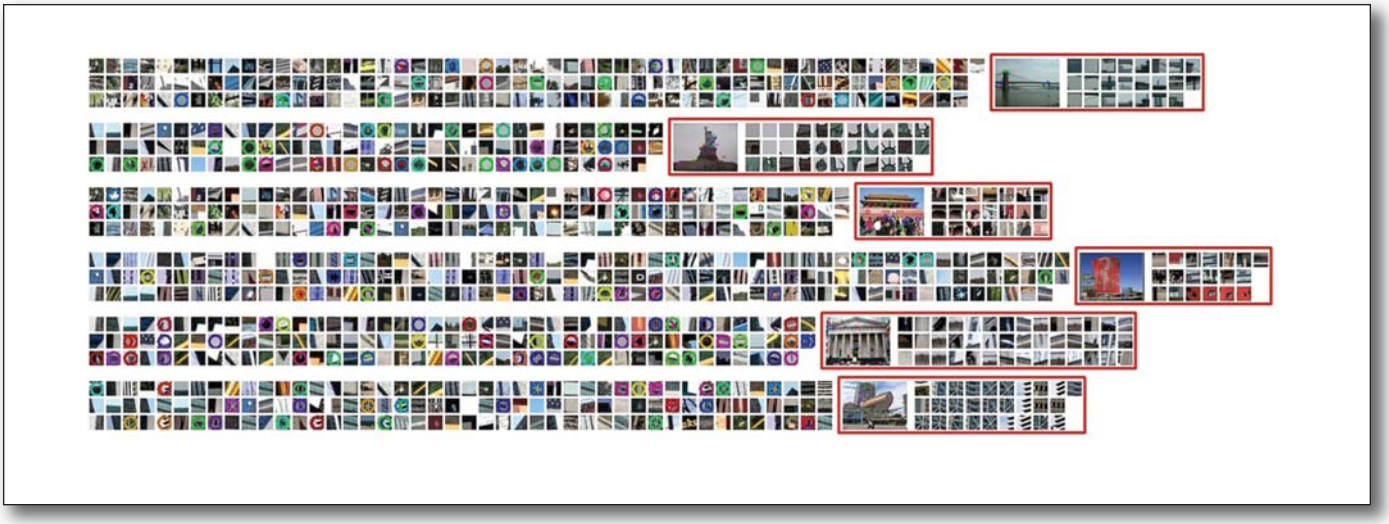

Figure 9. The Learned Compact Code Book and the Extracted Descriptors for Exemplar Queries in Barcelona.

Left: the compact code books in the query region. Middle: The query image, where color circles highlight the detected descriptors. *Right:* The actually transmitted words. We just transmit their occurrence index in practice.

with adaptive and error-resistant channel selection, can largely boost the performance of visual search. The proposed MCVD descriptor further exploits rich contextual cues available at the mobile end (such as GPS, two-dimensional barcodes, or RFID tags), as well as the visual feature statistics of the reference image database, to learn compact descriptors for mobile visual search rather than just landmark search. Undoubtedly, the readily available mobile context can significantly improve the visual search performance, which has been evidenced in papers by Ji et al. (2011c) and Ji et al. (2012).

Discussions Under the Umbrella of Standardization

Academe and industry have made progress on key technical components for visual search (figure 3); however, a few practical issues remain. It is unclear, for example, how to make visual search applications compatible across a broad range of devices and platforms. In this section, we extend our landmark search to visual descriptor standardization. Comparisons between landmark search and generic visual search will be discussed as well.

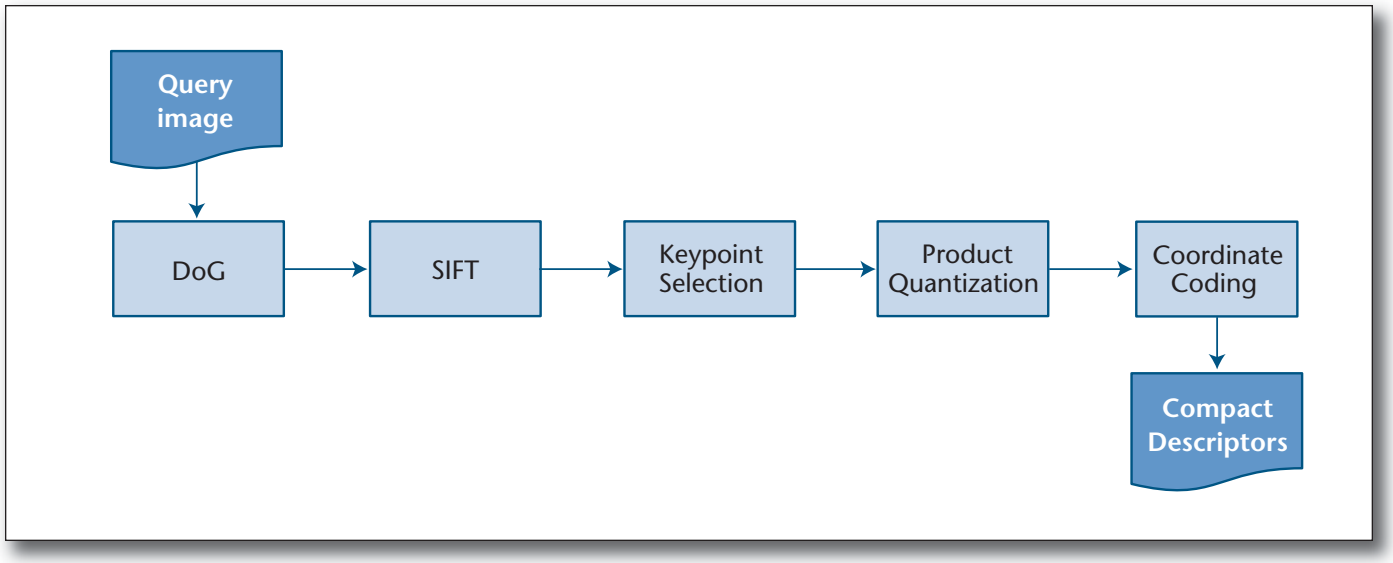

Figure 10. The CDVS Test Model Under Consideration.

Two beginning blocks extract raw local descriptors including a multiscale key-point detector based on difference of Gaussian (DoG) and a SIFT descriptor based on gradient-orientation histograms. Key-point selection filters in a subset of important key points to fulfill the compactness and scalability, subject to bandwidth budgets. Product quantization compresses each SIFT descriptor in a data-driven manner, and coordinate coding is applied to compress key-point locations. The test model under consideration represents the latest compact descriptor in the ongoing MPEG standardization.

Meeting	Time	Milestone
97th	Jul-2011	Call for proposal issued
98th	Dec-2011	Proposal evaluated, Test model
99th	Feb-2012	Six core experiments set up
100th	Apr-2012	Evaluation of proposals
101st	Jul-2012	First working draft
103rd	Jan-2013	Committee draft
105th	Jul-2013	Draft International Standard
107th	Feb-2014	Final Draft International Standard

Table 1. Timeline of MPEG CDVS Standarization.

Background of MPEG CDVS

To ensure interoperability, the MPEG compact descriptor for visual search standardization aims to define the format of compact visual descriptors. Readers are referred to table 1 for MPEG CDVS standardizations timeline and milestones.[6]

In terms of CDVS requirements (Yuri et al. 2011), visual descriptors shall be robust, compact, and easy to compute on a wide range of platforms. High matching accuracy shall be achieved at least for images of rigid, textured objects, landmarks, and documents. Matching should be accurate despite partial occlusions and changes in vantage point, camera parameters, and lighting. To reduce query transmission latency, the descriptor length shall be minimized. Adaptation of descriptor length is enabled so that the performance level can be satisfied at expected budgets. Extracting descriptors cannot be too complex in terms of memory and time.

Evidences from Low Bit Rate Landmark Search

To determine the lowest operating point for promising visual search,[7] geotag (a kind of side information) has been used to produce very compact descriptors for visual search of landmarks (Ji et al. 2012). One important budget evidence is what this article reported on low bit rate landmark search, namely, location-discriminative vocabulary coding (LDVC). With hundreds of bits per query image,[8] LDVC encodes descriptors over quantized SIFT features (Lowe 2004).

As mentioned previously, beyond landmark search, visual statistics and mobile context have been jointly exploited over generic image databases to come up with multiple coding channels (Ji et al. 2011d). A compression function is learned for each channel. Each query is initially represented by a high-dimensional visual signature, which is then mapped to one or more channels for further compression. Likewise, with just hundreds of bits, a compact descriptor has achieved comparable promising search to raw SIFT features.

Based on the evidence of hundreds of bits in yielding a compact descriptor, the MPEG CDVS ad hoc group has finally determined the lowest oper-

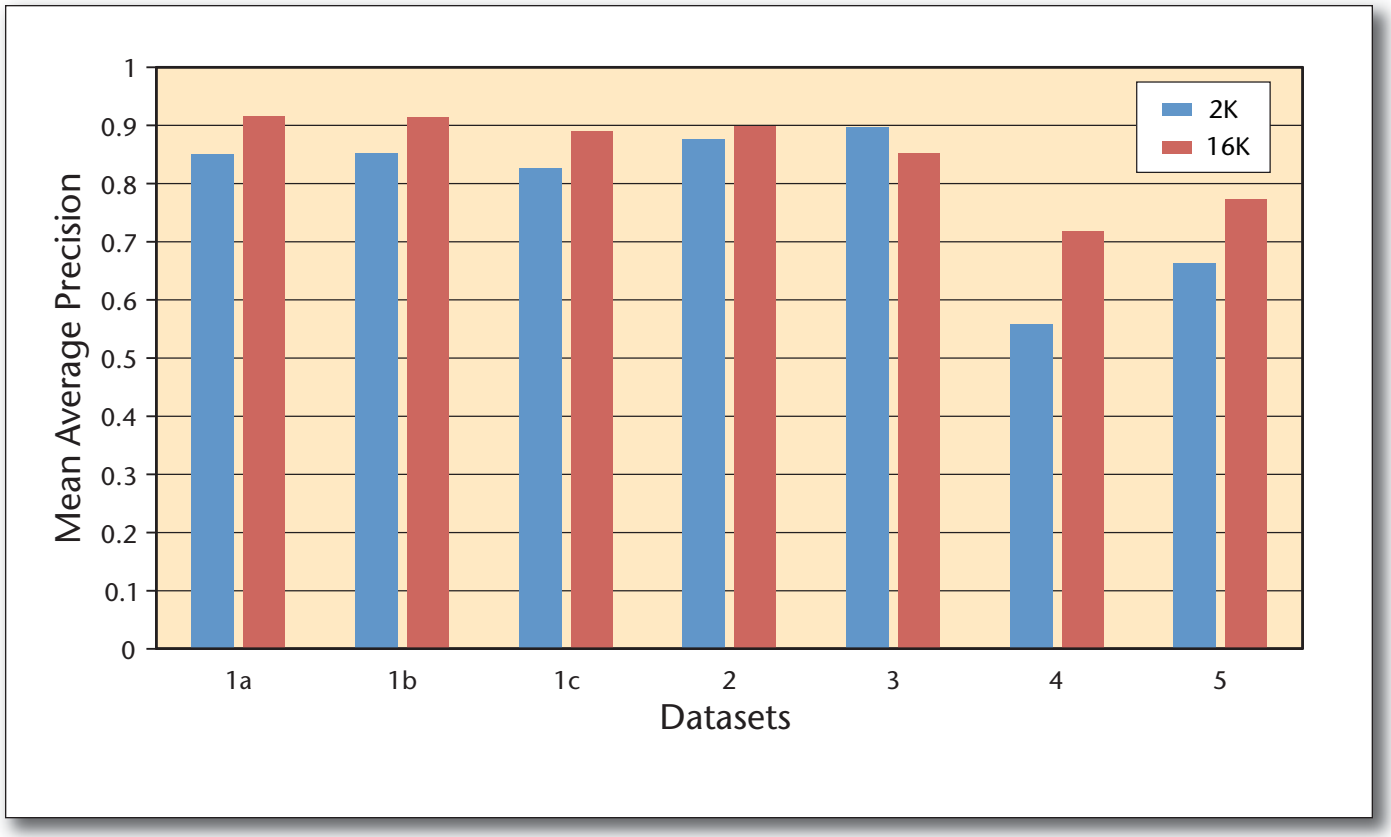

Figure 11. Exemplar Query and Reference Images of Different Categories in the CDVS Evaluation Data Set.

(a) Mixed (CD/DVD covers, book covers, documents), (b) paintings, (c) video frames, (d) landmarks, (e) common objects.

ating point of 512 bytes (including the coordinates of local descriptors).

Compact Landmark Descriptors Versus the Test Model Under Consideration

Based on the responses to the CDVS call for papers issued at the Torino meeting,[9] a test model under consideration based on product quantization (Gray and Neuhoff, 1998) was selected at the 98th MPEG Geneva meeting (Francini et al. 2011). We now discuss the link between compact landmark descriptors and the test model under consideration.

Figure 10 shows the flowchart of the test model under consideration. In this model, with vector quantization (VQ), descriptor size can be reduced significantly; for example, in the test model (Francini et al. 2011), more than 85 percent of bits are saved. In addition, feature locations are compressed by quantization and context arithmetic coding.

To reduce the encoding complexity (normally increasing dramatically with the vector dimension), the test model under consideration employs product quantization to divide an input vector into k segments and quantize those segments independently using k subquantizers. It is easy to imagine, when $k = 1$, the test model under consideration's visual descriptor has degenerated to the basic pipeline of compact landmark descriptor (Ji et al. 2011b).

Moreover, reducing the quantization complexity is meaningful for mobile applications. Alternatively, our proposed boosting-based code-book learning in compact landmark descriptors can shrink the code-word scale of tree-structured vector quantizer (TSVQ) (that is, the SVT model) in Chen et al. (2012), where memory cost is reduced from over 60 MB to below 10 MB.[10] In the test model under consideration, the product quantization table's memory cost can be reduced to 128 KB (Chen et al. 2012).

Landmark Search Versus Visual Search

Figure 11 lists exemplar images of different categories in MPEG CDVS evaluation framework (Yuri et al. 2011). Compared to planar objects (CD/DVD covers, book covers, paintings, and others), searching three-dimensional landmarks is the most challenging. The lowest mean average precision in figure 12 provides quantitative evidences (Li et al.

Figure 12. Retrieval Results at Two Operating Points of 2 K and 16 K on the CDVS Evaluation Data Sets.

(For example, the descriptor length per query image). (1a)(1b)(1c) mixed, (2) paintings, (3) video frames, (4) landmarks, (5) common objects, referring to exemplar images in figure 12. Experiments are performed over 30,256 reference images plus 1 million distractor images (collected from Flickr). The mean average precision comparison indicates that landmark search is the most challenging task.

2012) on the challenges of landmark search, compared with other planar objects.

Undoubtedly, landmarks represent a typical kind of nonplanar object, often incurring ill-posed two-dimensional photographic configurations and variances in occlusion, viewpoint, scale, illumination, and background. Concrete challenges include clutter (for example, vehicles, pedestrians, seasonal vegetation), shadows cast on buildings, reflections and glare on windows, and severe perspectives with extreme angles. Large photometric and geometric distortions separate query images from their closest matches in the reference databases. So context-assisted descriptors are practically useful.

These aforementioned challenges justify why our visual search research focuses on landmark objects. Beyond the challenges of landmark search, we believe that the insights of contextual learning as well as ranking of sensitive vocabulary boosting to achieve low bit rate descriptors can also be extended to generic visual search.

Conclusions

In this article, we have proposed to learn a compact visual descriptor by combining both visual content and geographical context; this technique has been deployed in real-world mobile landmark search applications. To fulfill extreme compactness for low bit rate query transmission, our proposed contextual learning came up with an iterative optimization scheme by combining geographical segmentation and descriptor learning. The compact descriptor has been deployed in both HTC Desire G7 (Android) and iPhone4 (iOS) platforms and significantly outperforms state-of-the-art works (Nistér and Stewénius 2006; Chen et al. 2009; Chandrasekhar et al. 2009a; Jegou et al. 2010) over a large data set of 1 million landmark images. In addition to desirable search accuracy and reduced query delivery latency, the mobile user experiences relates to factors such as the user interface, value-added information, and others. Comprehensive user study will be included in our future work.

With the progress of MPEG standardization, we envision the use of compact visual descriptors in a wide range of visual search applications including mobile augmented reality. More importantly, the ongoing standardization has attracted the interest of hardware manufacturers such as Aptina, Nvidia, STMicroelectronics, and others. The proposed context-assisted compact descriptor as well as the zero-latency visual search provided strong research evidence in determining operating points and choosing vector quantization to compress local descriptors. As landmark search is the most challenging, the compact landmark descriptor may be

considered as a good reference model of generic compact descriptors. However, context is not included into CDVS core experiments (Yuri et al. 2012) due to the challenging issue of collecting context-tagged data sets covering rich objects in addition to the geotagged PKUBench data set (Ji et al. 2011c).

Acknowledgements

This work was supported by the Chinese National Natural Science Foundation under contracts No. 60902057 and No. 61271311, and in part by the National Basic Research Program of China under contract No. 2009CB320902.

Notes

1. Note that the contextual learning consists of two iterative stages: (1) optimizing the partition of geotagged reference images; and (2) learning the compact vocabulary within each partition. The context learning is performed offline. In online search, a mobile phone just exploits the location context to determine the partition and accordingly select its corresponding compact vocabulary for generating compact descriptors. The vocabularies may be prestored in a mobile phone or downloaded at the mobile users' request. In other words, a mobile phone does not involve intensive online computation from context learning.

2. Note that our landmark search work does not use additional context beyond location, although context information often provides rich meanings or functionalities.

3. For the convenience of explanation, we denote scalars as italic letters, for example, v; vectors as bold letters, for example, $\mathbf{v}$; instance spaces for n as R_n; and the inner product between u and v as

$$\mathbf{u} \cdot \mathbf{v} = \sum_{i=1}^{n} u_i v_i$$

4. Cosine similarity is a measure of similarity between two vectors by measuring the cosine of the angle between them. This is often used to compare documents in text retrieval. The resulting similarity ranges from –1, meaning exactly opposite, to 1, meaning exactly the same, with 0 usually indicating independence and in between values indicating intermediate similarity or dissimilarity.

5. For a more practical viewpoint, dense regions are selected from popular tourism zones. The size of location regions mainly depends on the photo volume, which usually ranges from 0.2 to 0.5 kilometer in diameter. For hot zones (like Tiananmen Square in Beijing city), the region diameter is less than 0.1 kilometer, while the diameter of suburb regions even reachs up to 5 kilometer.

6. See M. Bober et al., Description of Core Experiments on Compact Descriptors for Visual Search (MPEG ISO/IEC JTC1/SC29/WG11/N12551, 2012/2).

7. To evaluate the descriptor scalability, MPEG CDVS requires proposals to report the results at 6 operating points with descriptor lengths per query: 512 bytes, 1 KB, 2 KB, 4 KB, 8 KB, and 16 KB.

8. Extra bits for location coding are not counted here.

9. See Yuri Reznik et al., Call for Proposals for Compact Descriptors for Visual Search (ISO/IEC JTC1/SC29/WG11 N12201, 2011/07); K. Iwamoto, R. Mase, et al., NEC's Response to CFP for Compact Descriptor for Visual Search (MPEG ISO/IEC JTC1/SC29/WG11/ M22717, 2011/11); V. Chandrasekhar, G. Kirsch, et al., CDVS Proposal: Stanford Nokia Aptina Features (MPEG ISO/IEC JTC1/SC29/WG11/ M22554, 2011/11); C. Wang, L.-Y. Duan, J. Chen, and R. Ji, Peking Compact Descriptor-PQ-WGLOH (MPEG ISO/IEC JTC1/SC29/WG11/M22619, 2011/11); and J. Chen, L.-Y. Duan, Rongrong Ji, et al., Peking Compact Descriptor: PQ-SIFT (MPEG ISO/IEC JTC1/SC29/WG11/M22620, 2011/11).

10. Smart phones have memory limit per process; for example, the limit is 16 MB for the first generation Android phone and 24 MB for the second generation. A smaller footprint is necessary for apps.

References

Bay, H.; Tuytelaars, T.; and Van Gool, L. 2006. SURF: Speeded Up Robust Features. In *Proceedings of the European Conference on Computer Vision,* Lecture Notes in Computer Science 3951, 404–417. Berlin: Springer.

Chandrasekhar, V.; Takacs, G.; Chen, D. M.; Tsai, S. S.; Grzeszczuk, R.; and Girod, B. 2009a. CHoG: Compressed Histogram of Gradients: A Low Bit-Rate Feature Descriptor. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* 2504–2511. Piscataway, NJ: Institute of Electrical and Electronics Engineers.

Chandrasekhar, V.; Takacs, G.; Chen, D.; Tsai, S. S.; Singh, J.; and Girod, B. 2009b. Transform Coding of Image Feature Descriptors. In *Proceedings of the SPIE Conference on Visual Communications and Image Processing*. Bellingham, WA: The International Society for Optical Engineering.

Chen, D. M.; Tsai, S. S; Chandrasekhar, V.; Takacs, G.; Singh, J. P.; and Girod, B. 2009. Tree Histogram Coding for Mobile Image Matching. In *Proceedings of the IEEE Data Compression Conference,* 143–152. Los Alamitos, CA: IEEE Computer Society.

Chen, D. M.; Tsai, S. S.; Chandrasekhar, V.; Takacs, G.; Vedantham, R.; Grzeszczuk, R.; and Girod, B. 2010. Inverted Index Compression for Scalable Image Matching. In *Proceedings of the IEEE Data Compression Conference,* 525. Los Alamitos, CA: IEEE Computer Society.

Chen, J.; Duan, L.-Y.; Ji, R.; and Gao, W. 2012. Pruning Tree-Structured Vector Quantizer Towards Low Bit Rate Mobile Visual Search. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*. Piscataway, NJ: Institute of Electrical and Electronics Engineers.

Chen, J.; Duan, L.-Y.; Ji, R.; Yao, H.; Gao, W. 2011. Sorting Local Descriptor for Low Bit Rate Mobile Visual Search. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing,* 1029–1032. Piscataway, NJ: Institute of Electrical and Electronics Engineers.

Crandall, D. J.; Backstrom, L.; Huttenlocher, D.; and Kleinberg, J. 2009. Mapping the World's Photos. In *Proceedings of Eighteenth International World Wide Web Conference,* 761–770. New York: Association for Computing Machinery.

Francini, G.; Lepsoy, S.; Balestri, M. 2011. Description of Test Model Under Consideration for CDVS. In Resolutions of the 98th Meeting of the Moving Picture Experts

Group N12254. Geneva, Switzerland: Moving Picture Experts Group.

Gemert, J.; Veenman, C.; Smeulders, A.; and Geusebroek, J. 2009. Visual Word Ambiguity. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 32(7): 1271–1283.

Girod, B.; Chandrasekhar, V.; Chen, D.; Cheung, N.-M.; Grzeszczuk, R.; Reznik, Y.; Takacs, G.; Tsai, S.; and Vedantham, R. 2011. Mobile Visual Search. *IEEE Signal Processing Magazine* 28(4): 61–76.

Gray, R. M., and Neuhoff, D. L. 1998. Quantization. *Information Theory* 44(6): 2325–2383.

Hays, J., and Efros, A. 2003. IMG2GPS: Estimating Geographic Information from a Single. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* 1–8. Piscataway, NJ: Institute of Electrical and Electronics Engineers.

Hua, G.; Brown, M.; and Winder, S. 2007. Discriminant Embedding for Local Image Descriptors. In *Proceedings of the IEEE International Conference on Computer Vision*, 1–8. Piscataway, NJ: Institute of Electrical and Electronics Engineers.

Irschara, A.; Zach, C.; Frahm, J.-M.; and Bischof, H. 2009. From SFM Point Clouds to Fast Location Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* 2599–2606. Piscataway, NJ: Institute of Electrical and Electronics Engineers.

Jegou, H.; Douze, M.; Schmid, C.; and Perez, P. 2010. Aggregating Local Descriptors into a Compact Image Representation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Piscataway, NJ: Institute of Electrical and Electronics Engineers.

Jegou, H.; Douze, M.; and Schmid, C. 2008. Hamming Embedding and Weak Geometric Consistency for Large Scale Image Search. In *Proceedings of the European Conference on Computer Vision,* Lecture Notes in Computer Science 5301, 304–317. Berlin: Springer.

Ji, R.; Duan, L.-Y.; Chen, J.; Yao, H.; and Gao, W. 2011a. A Low Bit Rate Vocabulary Coding Scheme for Mobile Landmark Search. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*. Piscataway, NJ: Institute of Electrical and Electronics Engineers.

Ji, R.; Duan, L.-Y.; Chen, J.; Yao, H.; Huang, T.; and Gao, W. 2011b. Learning Compact Visual Descriptor for Low Bit Rate Mobile Landmark Search. In *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence,* 2456–2463. Menlo Park, CA: AAAI Press.

Ji, R.; Duan, L.-Y.; Chen, J.; Yao, H.; Huang, T.; Yao, H.; and Gao, W. 2011c. PKUBench: A Context Rich Mobile Visual Search Benchmark. In *Proceedings of the 18th IEEE International Conference on Image Processing,* 2545–2548. Piscataway, NJ: Institute of Electrical and Electronics Engineers.

Ji, R.; Duan, L.-Y.; Chen, J.; Yao, H.; Rui, Y.; Chang, S.-F.; and Gao, W. 2011d. Towards Low Bit Rate Mobile Visual Search with Multiple Channel Coding. In *Proceedings of the 19th ACM International Conference on Multimedia,* 573–582. New York: Association for Computing Machinery.

Ji, R.; Duan, L.-Y.; Chen, J.; Yao, H.; Yuan, J.; Rui, Y.; and Gao, W. 2012. Location Discriminative Vocabulary Coding for Mobile Landmark Search. *International Journal of Computer Vision* 96(3): 290–314.

Ji, R.; Yao, H.; Sun, X.; and Gao, W. 2010. Towards Semantic Embedding in Visual Vocabulary. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* 2504–2511. Piscataway, NJ: Institute of Electrical and Electronics Engineers.

Ji, R.; Xie, X.; Yao, H.; and Ma, W.-Y. 2009. Hierarchical Optimization of Visual Vocabulary for Effective and Transferable Retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Piscataway, NJ: Instite of Electrical and Electronics Engineers.

Jiang, Y.-G.; Ngo, C.-W.; and Yang, J. 2007. Towards Optimal Bag-of-Features for Object Categorization and Semantic Video Retrieval. In *Proceedings of the 6th ACM International Conference on Image and Video Retrieval,* 494–501. New York: Association for Computing Machinery.

Jurie, F., and Triggs, B. 2005. Creating Efficient Codebooks for Visual Recognition. In *Proceedings of the IEEE International Conference on Computer Vision,* 604–610. Los Alamitos, CA: IEEE Computer Society.

Ke, Y., and Sukthankar, R. 2004. PCA-SIFT: A More Distinctive Representation for Local Image Descriptors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* 506–513. Piscataway, NJ: Institute of Electrical and Electronics Engineers.

Kulis, B., and Grauman, K. 2009. Kernelized Locality-Sensitive Hashing for Scalable Image Search, 2130–2137. In *Proceedings of the 12th IEEE International Conference on Computer Vision*. Piscataway, NJ: Institute of Electrical and Electronics Engineers.

Lazebnik, S., and Raginsky, M. 2009. Supervised Learning of Quantizer Codebooks by Information Loss Minimization. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 31(7): 1294–1309.

Li, B.; Duan, L.-Y.; Chen, Y.; Lin, J.; Huang, T.; Gao, W. 2012. Improvements for the Retrieval Pipeline of TM2.0. In Resolutions of the 101st Meeting of the Moving Pictures Experts Group (MPEG) M25904. Stockholm, Sweden: Moving Pictures Experts Group.

Li, X.; Wu, C.; Lazebnik, S.; and Frahm, J.-M. 2008. Modeling and Recognition of Landmark Image Collections Using Iconic Scene Graphs. In *Proceedings of the European Conference on Computer Vision,* Lecture Notes in Computer Science, 427–440. Berlin: Springer.

Liu, D.; Scott, M.; Ji, R.; Yao, H; and Xie, X. 2009. Location Sensitive Indexing for Image-Based Advertising. In *Proceedings of the 17th ACM International Conference on Multimedia,* 793–796. New York: Association for Computing Machinery.

Lowe, D. G. 2004. Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision* 60(2): 91–110.

Mairal, J.; Bach, F. ; Ponce, J.; Sapiro, G.; and Zisserman, A. 2008. Supervised Dictionary Learning. In *Proceedings of the Neural Information Processing Systems Conference*. Cambridge, MA: The MIT Press.

Makar, M.; Chang, C.; Chen, D.; Tsai, S.; and Girod, B. 2009. Compression of Image Patches for Local Feature Extraction. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing,* 821–824. Piscataway, NJ: Institute of Electrical and Electronics Engineers.

Mikolajczyk, K., and Schmid, C. 2005. Performance Evaluation of Local Descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* (27)10: 1615–1630.

Moosmann, F.; Triggs, B.; and Jurie, F. 2006. Fast Discriminative Visual Codebooks Using Randomized Clustering Forests. In *Proceedings of the Neural Information Processing Systems Conference*. Cambridge, MA: The MIT Press.

Nistér, D., and Stewénius, H. 2006. Scalable Recognition with a Vocabulary Tree. 2006. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* 2161–2168. Piscataway, NJ: Institute of Electrical and Electronics Engineers.

Philbin, J.; Chum, O.; Isard, M.; Sivic, J.; and Zisserman, A. 2007. Object Retrieval with Large Vocabulary and Fast Spatial Matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Piscataway, NJ: Institute of Electrical and Electronics Engineers.

Salton, G., and Buckley, C. 1988. Term-Weighting Approaches in Text Retrieval. *Information Processing and Management: An International Journal* 24(5): 513–523.

Schindler, G.; Brown, M.; Szeliski, R. 2007. City-Scale Location Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* 2504–2511. Piscataway, NJ: Institute of Electrical and Electronics Engineers.

Shao, H.; Svoboda, T; Tuytelaars, T.; Van Gool, L. 2003. HPAT Indexing for Fast Object/Scene Recognition Based on Local Appearance. In *Image and Video Retrieval: Proceedings of the Second International Conference,* 71–80. Berlin: Springer.

Sivic, J., and Zisserman, A. 2003. Video Google: A Text Retrieval Approach to Object Matching in Videos. In *Proceedings of the IEEE International Conference on Computer Vision,* 1470–1477. Piscataway, NJ: Institute of Electrical and Electronics Engineers

Stauffer, C., and Grimson, W. 2000. Learning Patterns of Activity Using Real-Time Tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22(8): 747–757.

Tsai, S. S.; Chen, D.; Takacs, G.; Chandrasekhar, V.; Singh, C. J.; and Girod, B. 2009. Location Coding for Mobile Image Retrieval. In *Proceedings of the 5th International Mobile Multimedia Communications Conference*. New York: Association for Computing Machinery.

Yuri R.; Miroslaw B.; Giovanni C. 2010. Compact Descriptors for Visual Search: Context and Objectives. In Document Archive of the 93rd Moving Picture Experts Group, N11531., Geneva, Switzerland: MPEG.

Yuri R.; Miroslaw B.; Giovanni C. 2011. Call for Proposals for Compact Descriptors for Visual Search. In Document Archive of the 97th Moving Picture Experts Group Meeting, N12201, Torino, Italy: MPEG.

Zhang, W., and Kosecka, J. 2006. Image Based Localization in Urban Environments. In *Proceedings of the IEEE International Symposium on 3D Data Processing, Visualization, and Transmission,* 33–40. Piscataway, NJ: Institute of Electrical and Electronics Engineers

Zheng, Y. T.; Zhao, M.; Song, Y.; and Adam, H. 2009. Tour the World: Building a Web-Scale Landmark Recognition Engine. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* 1085–1092. Piscataway, NJ: Instite of Electrical and Electronics Engineers.

Ling-Yu Duan received a Ph.D in information technology from the University of Newcastle, Australia, in 2007, an M.Sc. degree in computer science from the National University of Singapore, Singapore, and an M.Sc. degree in automation from the University of Science and Technology of China, Hefei, China, in 2002 and 1999, respectively. Since 2008, he has been with Peking University, Beijing, China, where he is currently an associate professor with the School of Electrical Engineering and Computer Science. Duan leads the visual search group in the Institute of Digital Media, Peking University. Before that, he was a research scientist in the Institute for Infocomm Research, Singapore. His interests are in the areas of visual search and augmented reality, multimedia content analysis, and mobile media computing. He has authored more than 80 publications in these areas.

Jie Chen is a Ph.D. candidate at the School of Electronics Engineering and Computer Science, Peking University. He works with Ling-Yu Duan in the Institute of Digital Media. His research topics include data compression and visual search, focusing on compact visual descriptors for large-scale mobile visual search.

Rongrong Ji is a postdoctoral researcher at the Electronic Engineering Department, Columbia University. He received his Ph.D. from the Computer Science Department, Harbin Institute of Technology. His research interests include image retrieval and annotation, video retrieval and understanding. He has been a research intern at Microsoft Research Asia, where he received a Microsoft Fellowship, and a visiting student at the Institute of Digital Media, Peking University. He has published papers in more than 50 referred journals and conferences.

Tiejun Huang received B.S. and M.S. degrees from the Department of Automation, Wuhan University of Technology, Wuhan, China, in 1992 and a Ph.D. degree from the School of Information Technology and Engineering, Huazhong University of Science and Technology, Wuhan, China, in 1999. He was a postdoctoral researcher from 1999 to 2001 and a research faculty member at the Institute of Computing Technology, Chinese Academy of Sciences. He is currently a professor at National Engineering Laboratory for Video Technology, School of Electrical Engineering and Computer Science, Peking University, China. His research interests include digital media technology, digital library, and digital rights management.

Wen Gao received a Ph.D. degree in electronics engineering from the University of Tokyo, Tokyo, Japan, in 1991. He is a professor of computer science with the Institute of Digital Media, School of Electronic Engineering and Computer Science, Peking University, Beijing, China. Before joining Peking University, he was a professor of computer science with the Harbin Institute of Technology, Harbin, China, and a professor with the Institute of Computing Technology, Chinese Academy of Sciences, Beijing. Gao served or serves on the editorial boards for several journals and has chaired a number of international conferences on multimedia and video signal processing.