Mining Compact Bag-of-Patterns for Low Bit Rate Mobile Visual Search

Rongrong Ji, *Senior Member, IEEE*, Ling-Yu Duan, *Member, IEEE*, Jie Chen, *Student Member, IEEE*, Tiejun Huang, *Member, IEEE*, and Wen Gao, *Fellow, IEEE*

***Abstract*— Visual patterns, i.e., high-order combinations of visual words, contributes to a discriminative abstraction of the high-dimensional bag-of-words image representation. However, the existing visual patterns are built upon the 2D photographic concurrences of visual words, which is ill-posed comparing with their real-world 3D concurrences, since the words from different objects or different depth might be incorrectly bound into an identical pattern. On the other hand, designing compact descriptors from the mined patterns is left open. To address both issues, in this paper, we propose a novel compact bag-of-patterns (CBoPs) descriptor with an application to low bit rate mobile landmark search. First, to overcome the ill-posed 2D photographic configuration, we build up a 3D point cloud from the reference images of each landmark, therefore more accurate pattern candidates can be extracted from the 3D concurrences of visual words. A novel gravity distance metric is then proposed to mine discriminative visual patterns. Second, we come up with compact image description by introducing a CBoPs descriptor. CBoP is figured out by sparse coding over the mined visual patterns, which maximally reconstructs the original bag-of-words histogram with a minimum coding length. We developed a low bit rate mobile landmark search prototype, in which CBoP descriptor is directly extracted and sent from the mobile end to reduce the query delivery latency. The CBoP performance is quantized in several large-scale benchmarks with comparisons to the state-of-the-art compact descriptors, topic features, and hashing descriptors. We have reported comparable accuracy to the million-scale bag-of-words histogram over the million scale visual words, with high descriptor compression rate (approximately 100-bits) than the state-of-the-art bag-of-words compression scheme.**

***Index Terms*— Mobile visual search, compact visual descriptor, low bit rate retrieval, bag-of-words, visual pattern mining, structure-from-motion, sparse coding.**

Manuscript received February 14, 2013; revised September 13, 2013 and February 10, 2014; accepted March 13, 2014. Date of publication May 14, 2014; date of current version June 11, 2014. This work was supported in part by the National Natural Science Foundation of China under Grant 61271311, 61390515, and 61210005, in part by the MDA Research Fund of the ROSE Laboratory, in part by the National Natural Science Foundation of China under Grant 61373076, in part by the Fundamental Research Funds for the Central Universities under Grant 2013121026, and in part by the 985 Project of Xiamen University, Xiamen, China. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Gang Hua. *(Corresponding author: Ling-Yu Duan.)*

R. Ji is with the Institute of Digital Media, School of Electronics Engineering and Computer Science, Peking University, Beijing 100871, China, and also with the School of Information Science and Technology, Xiamen University, Xiamen 361005, China.

L.-Y. Duan, J. Chen, T. Huang, and W. Gao are with the Institute of Digital Media, School of Electronics Engineering and Computer Science, Peking University, Beijing 100871, China (e-mail: lingyu@pku.edu.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2014.2324291

I. Introduction

MOST existing scalable visual search systems are built based on visual vocabulary with inverted indexing [1]–[4]. Local features extracted from the reference image are quantized into visual words, producing the so-called bag-of-words histogram and the image is inverted indexed in every non-zero words correspondingly. The bag-of-words representation offers sufficient robustness against photographing variances in occlusions, viewpoints, illuminations, scales and backgrounds. Then, image search can be addressed in a similar way as document retrieval, by which well-used techniques like TF-IDF [5], pLSA [6] and LDA [7] can be deployed.

Motivation. One significant drawback of bag-of-words comes from ignoring the spatial layouts of words. To a certain degree, this can be compensated by carrying out spatial post-verification, using for instance RANdom SAmple Consensus (RANSAC) or neighborhood voting [1]. Alternatively, a more efficient solution is to use pre-defined spatial coding to group spatially nearby words as new features, for instance feature bundling [8] and max/min pooling [9].

Nevertheless, the discriminability of pairing or grouping words in each image is *not alone*, which in turn highly depends on the overall statistics of word combinations in the rest images. To this end, rather than fixed spatial coding [8], [9], a more *data driven* approach is to discover such discriminative grouping of visual words from the image corpus. This is referred to as "visual patterns" or "visual phrases" [10]–[13] as from the information retrieval literatures, which typically involves techniques like co-location mining [14].

Formally speaking, a visual pattern is a meaningful spatial combination of visual words, which can be regarded as a semi-fully geometrical dependency feature, where the geometry of each part depends only on its neighbors. Comparing to previous works in spatial modeling where restrictive priors and parameters are demanded [15], [16], visual patterns have been well advocated by their parameter-free intrinsic, *i.e.* the pattern structures are obtained by data mining with class or category supervision. It equips visual patterns with the scalability, as advanced from the previous unscalable models [15], [16].

Problem. Two important problems are left open in the existing visual pattern mining paradigm:

- The existing visual patterns are built upon the 2D word concurrences in individual images. Such concurrence suffers from the ill-posed 2D photographic constitution and cannot capture the real-world 3D layouts of words. For instance words from different depth or different

Fig. 1. Exemplar illustrations of incorrect 2D neighborhood configurations of visual words, which are caused by either binding words with diverse depth, or binding words from both foreground and background objects.

foreground/background objects may be encoded together from the 2D perspective. However, such spatial concurrence and coding are not robust and discriminative, as words may be disparity from other perspective. Figure 1 shows several examples to explain such incorrect configurations.

- Given the mined patterns, how to design a compact yet discriminative image representation is left open. To the best of our knowledge, visual patterns are typically concatenated with the bag-of-words histogram [10]–[13], as similar to the usage of textual patterns in the document retrieval endeavor. We argue that, given a well-designed pattern selection strategy, visual pattern alone is already discriminative enough[1]. Our explanation is that the visual word dependency has a clear spatial structure, which is more discriminative comparing to the contextual concurrence of textual words. This pattern-level compact descriptor well suits for several emerging applications like low bit rate mobile visual search [17], as detailed in Section V.

Approach. We propose a Compact Bag-of-Patterns (CBoP) descriptor to address both issues towards a compact yet discriminative image representation. Figure 2 outlines the work flow of our CBoP descriptor, which is built over the popular bag-of-words model. In preliminary, we assume that each target (*e.g.* object or landmark) in the dataset contains multiple reference images captured at different viewing angles. Using these images, a 3D point cloud is built for this target by structure-from-motion [18]. Then, we present a 3D sphere coding scheme to construct the initial pattern candidates, which eliminates the ill-posed 2D spatial layout in individual images by binding 3D concurrent visual words in the point cloud.

Next, in visual pattern mining, we introduce a "*gravity distance*" to incorporate the relative importance statistics of words, i.e., the mutual information between their frequencies, into the subsequent Aprior based frequent itemset mining [19]. In such a manner, the pattern reflects both concurrence frequency in spatial space and statistical saliency in feature space.

Finally, a Compact Bag-of-Patterns (CBoP) is built from the mined patterns by *pooling*. Pooling seeks an optimal tradeoff between the descriptor compactness and its discriminability. Sparse coding is employed to minimize the number of selected patterns (typically at hundreds of bits) under a given distortion tolerance between CBoP and the originally bag-of-words. It is worth to note that, supervised labels can be also integrated into the above formulation to further improve the performance.

Applications. The mined CBoP descriptor has potentials in multidisciplinary applications such as object recognition, visual search and image classification. In this paper, we demonstrate its usage in the *low bit rate mobile visual search* application, where visual descriptors are directly extracted on the mobile device and sent instead of the query image to reduce the query delivery latency [17], [20], [21]. In such a scenario, the extracted descriptor is expected to be compact and discriminative, and the extraction should be efficient. Different from the state-of-the-art works of directly compressing the high-dimensional bag-of-words histogram [17], [20], the pattern-level abstraction is a more natural choice yet left open so far. We provide two additional arguments to support such patter-level descriptors:

- First, previous works are deployed based on the linear combination of visual words for instance boosting [17], which selects one word in each round into the compact descriptor. It is a natural extension to look at their higher-order statistical combinations, e.g. patterns, to further improve the compression rate.
- Second, we argue that a pattern-level descriptor benefits in both memory cost and extraction time. In this case, only linear operations are applied to the initial bag-of-words, which is memory light and much faster comparing to alternatives like topic features [22], [23], [49].

In practice, our CBoP has achieved almost identical search accuracy comparing to the high-dimensional (million scale) bag-of-words histogram, with an approximate 100-bit descriptor size. It significantly outperforms state-of-the-art alternative descriptors like 2D visual patterns [12], [13], topic features [22], [23], [49] and Hashing based descriptors [24].

Outline. The rest of this paper is organized as follows: Section II reviews related work in visual vocabulary, visual pattern mining and compact descriptor etc. Section III introduces our discriminative 3D visual pattern mining and CBoP extraction schemes. Section V shows its application in low bit rate mobile landmark search, with quantitative comparisons to the state-of-the-art works [2], [12], [13], [22], [23], [24].

II. Related Work

Visual Vocabulary. Building visual vocabulary usually resorts to unsupervised vector quantization techniques such as k-means clustering [1], hierarchical k-means clustering (Vocabulary Tree) [2], approximate k-means clustering [3] and their variations [3], [25], [26]. Using visual vocabulary, an image is represented as a bag-of-words histogram, where each bin counts how many local features of this image are quantized into the corresponding word. To compensate the word uncertainty and ambiguity caused, the quantization strategy in bag-of-words can be further improved using methods such as Hamming Embedding [29], Soft Assignments [3], and kernelized visual codebook [30]. In general, visual vocabulary can be treated as an approximated nearest neighborhood search. From this perspective, an alternative solution is hashing, for instance Locality Sensitive Hashing (LSH) [27], kernalized LSH [28] and Spectral Hashing [24], which hashes each local feature into a few bins and indices the images accordingly.

[1]We show in Section III-B the selection strategy.

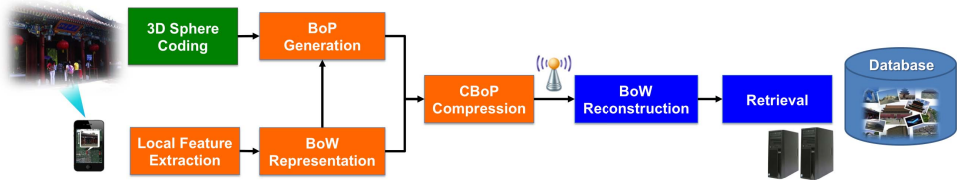

Fig. 2. The proposed Compact Bag of Patterns (CBoP) descriptor with an application to low bit rate mobile visual search.

Visual Pattern Mining. To further represent higher-order discriminability, one solution is to model the spatial distributions among words. For instance, Csurka *et al.* [31] proposed a generative model to characterize the spatial joint distribution of words. Crandall *et al.* [32] proposed a middle-level image representation by combining geometrically nearby words, which requires pre-defined model priors and supervised labels. However, these works are restricted due to the requirement of model parameters or priors, which is therefore unsuitable for the scenario of large-scale visual search.

A more promising alternative is to automatically discover meaningful high-order combinations, or so-called "patterns" or "phrases" [10]–[13]. Comparing to [31] and [32] with pre-defined model priors, a visual pattern can be treated as a lossy spatial representation that is discovered by unsupervised data mining. For example, works in [10]–[13] encode the spatial concurrence between words into item transactions, and then adopts co-location pattern mining [14] to discover patterns that are spatially nearby and frequently concurrent. Yuan et al. [12] proposed to mine class-specific patterns, which are in turn used to refine the local feature detection, e.g. to filter out unreliable points.

On the other hand, works in [54]–[56] proposed to directly combine the spatial layouts into the bag-of-words (BoW) construction (or more specially on the construction of contextual dictionary), which performs comparable to the visual pattern mining schemes. However, these papers focus on mining 2D patterns, which might be affected by the incorrect 2D photographic concurrence between visual words. In turn, working on 3D patterns serves as our key innovation. Extracting 3D patterns brings about more advantages compared to using 2D patterns only. View-based 3D object matching [53] has been investigated for many years, where the hypergraph-based method in Gao et al. [57] has shown the state-of-the-art performance. The results in existing 3D object retrieval has shown the superority of using multiviews on object representation.

Word Abstraction. Different from visual patterns, works in [33]–[36] focus on abstracting the initial words into a more compact representation in a supervised manner. For instance, Perronnin et al. [33] integrated category labels to adapt an initial vocabulary into several class-specific vocabularies. Winn et al. [36] learned class-specific vocabularies from an initial vocabulary by merging word pairs, in which the word distribution is modeled by the Gaussian mixture model. However, the requirement of category labels restricts the scalability of [33]–[36], especially given thousands categories or more. To a certain degree, topic models like probabilistic Latent Semantic Analysis [22] and Latent Dirichlet Allocation [23] can be also treated as a generative, higher-level word abstraction. However, super parameters are involved to control the topic generation, which is also time consuming and unscalable. We will quantitatively compare our CBoP descriptor to the state-of-the-art topic descriptor [49] in Section V.

Compact Descriptor for Mobile Visual Search. With the ever growing computation power of mobile devices, recent works [20], [21], and [17] have proposed to directly extract and send image descriptors on the mobile end to achieve a low cost wireless transmission. To this end, local descriptors in the literature like SIFT [37], SURF [38] or PCA-SIFT [39] are "over sized". For example, sending hundreds of such descriptors per image typically costs more data throughput comparing to sending the original image. One pioneer work in [21] proposed an extremely compact local descriptor named CHoG (Compressed Histogram of Gradients), which adopts tree coding to compress the initial descriptor into approximately 60 bits. In such a case, assuming the local detector outputs ~1,000 interest points per image, the overall data transmission is approximately 8KB, much less than sending the query image (typically over 20KB). Recent works in [20] and [17] stepped forward to directly compress the quantized bag-of-words histogram instead of the local descriptor set. For instance, Chen et al. [20] proposed to encode the position differences of the sparse non-zero histogram bins, resulting in an $\sim 2KB$ code per image for a vocabulary with 1 million words, which is much more compact than CHoG. Ji et al. [17] adopted boosting based supervised codeword selection to build location-adaptive codebooks for mobile landmark search, which achieves $\sim$ 60-bit descriptor per image with the help of side information such as GPS tags or landmark labels.

Location Recognition. There has been a longstanding history in vision-based location recognition. For example, for city-scale landmark recognition, Schindler et al. [4] presented a location recognition system through geo-tagged video streams with multiple-path search in the vocabulary tree [2]. Eade et al. [40] leveraged the vocabulary tree [2] based visual matching for real-time loop closing. Irschara et al. [41] adopted structure-from-motion to build 3D scene models for street views, which also combined vocabulary tree [2] based matching for simultaneously scene modeling and location recognition. Xiao et al. [50] proposed to combine bag-of-features with simultaneous localization and mapping to further improve the recognition precision. And incrementally vocabulary indexing is also explored in [51] to maintain a landmark search system in a time varying database.

For worldwide landmark recognition, the IM2GPS system [42] attempted to infer the possible location distributions of a given query by visually matching geo-tagged landmarks.

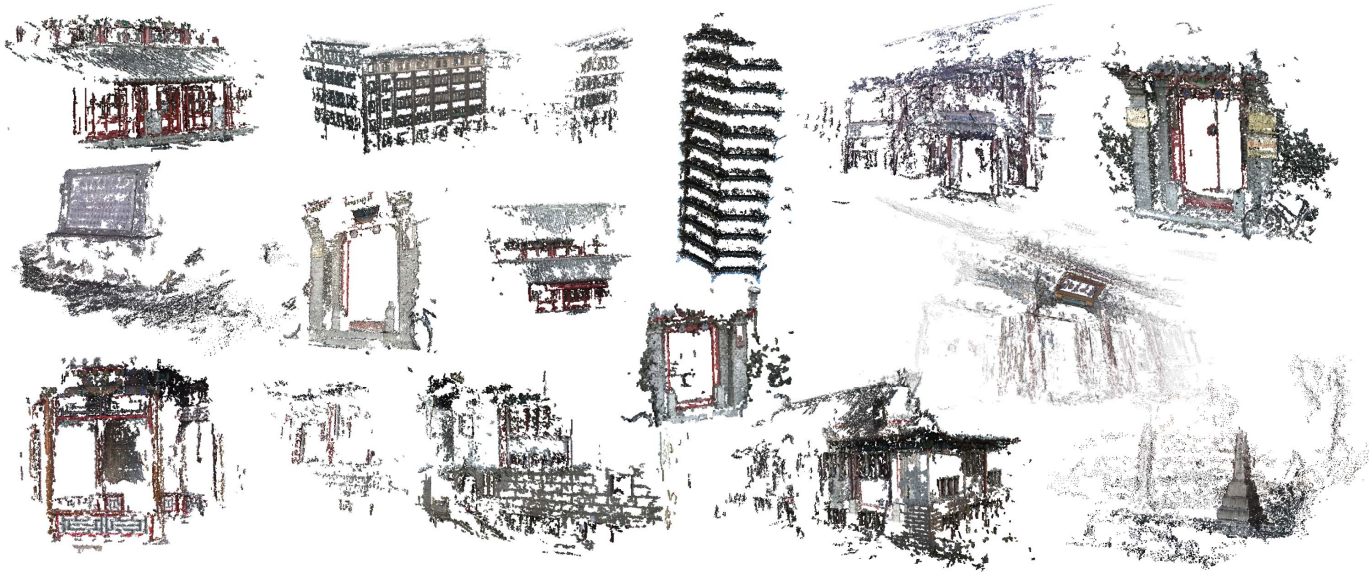

Fig. 3. Visualized examples of the point clouds for constructing visual pattern candidates. Exemplar landmarks are located at Peking Univ. Campus.

Kalogerakis et al. [43] further combined the single image matching with sequential data to improve the recognition accuracy. Zheng et al. [44] used a predefined landmark list to crawl candidate images from image search engines, then re-clusters and prunes these images to build a worldwide landmark recognition engine.

III. Discriminative 3D Pattern Mining

In this section, the 3D pattern mining and CBoP descriptor design is presented. We deploy CBoP based on the bag-of-words image representation as detailed in Section III-A. Suppose we have M Target of Interest (ToI) in total, each of which could be an object instance, a scene, a landmark or a product. For each ToI, we also have a set of reference images captured from different viewing angles. For generality, we assume the viewing angle parameters, i.e. inter and intra camera parameters, are unknown[2]. For each ToI, we first introduce a 3D sphere coding scheme to generate the initial pattern candidates, followed by a gravity-distance-based pattern mining algorithm in Section III-B. Finally, patterns from individual ToIs are pooled together to extract the Compact Bag-of-Patterns (CBoP) as detailed in Section III-C.

A. Scalable Visual Search Preliminary

Following [4], [17], [20], and [41], we use Vocabulary Tree (VT) [2] to build up the initial bag-of-words. VT adopts hierarchical k-means to partition local descriptors extracted from the image corpus into visual words. An h-depth VT with b-branch produces $m = b^h$ words. In many state-of-the-art systems m is typically at a million scale. Given a query image I_q with J local descriptors $\mathbf{L}(q) = [L_1(q), \ldots, L_J(q)]$, VT quantizes $\mathbf{L}(q)$ by traversing from the root to a leaf of the tree to find the nearest word for each descriptor. Therefore $\mathbf{L}(q)$ is converted into an m-dimensional BoW histogram $\mathbf{V}(q) = [V_1(q), \ldots, V_m(q)]$. Given in total N reference image $\{I_i\}_{i=1}^N$, an optimal ranking should minimize

$$Loss_{Rank}(q) = \sum_{i=1}^{N} Rank(i)\mathbf{W}_i\big(\mathbf{V}(q) - \mathbf{V}(i)\big) \tag{1}$$

where $Rank(i) = exp\big(-position(i)\big)$ is the ranking position of I_i and $\mathbf{W}_i$ is the TF-IDF vector for $\mathbf{V}(i)$ [5]:

$$\mathbf{W}_i = [\frac{N_{V_1 I_i}}{N_{I_i}} \times log(\frac{N}{N_{V_1}}), \ldots, \frac{N_{V_m I_i}}{N_{I_i}} \times log(\frac{N}{N_{V_m}})] \tag{2}$$

where N_i is the number of local descriptors in I_i; $N_{V_j I_i}$ the number of local descriptors in I_i quantized into V_j; N the total number of images in the database; N_{V_j} the number of images containing word V_j; $\frac{N_{V_j I_i}}{N_{I_i}}$ the term frequency of word V_j in image I_i; and $log(\frac{N}{N_{V_j}})$ the inverted document frequency of word V_j.

B. Discriminative 3D Pattern Mining

3D Sphere Coding. For each reference image I_i ($i \in [1, n_t]$) of the tth ToI ($t \in [1, M]$), suppose there are J local descriptors $\mathbf{L}(i) = [L_1(i), \ldots, L_J(i)]$ extracted from I_i, which is quantized into an m-dimensional bag-of-words histogram $\mathbf{V}(i) = [V_1(i), \ldots, V_m(i)]$. We denote the spatial positions of $\mathbf{L}(i)$ as $\mathbf{S}(i) = [S_1(i), \ldots, S_J(i)]$, where each $S_j(i)$ ($j \in [1, J]$) is the 2D or 3D spatial position of the jth local descriptor. For each $S_j(i)$, we scan its spatial k-nearest neighborhood to identify all concurrent words

$$\mathbf{T}_j(I_i) = \Big\{L_{j'}(i) | L_{j'}(i) \in \mathbf{L}(i) \;\&\; S_{j'}(i) \in kNN\big(S_j(i)\big)\Big\} \tag{3}$$

where $\mathbf{T}_j(I_i)$ (if any) is called an "transaction". We denote all transactions in I_i with order K as:

$$\mathbf{T}^k(I_i) = \big\{\mathbf{T}_j(I_i)\big\}_{j \in [1, J]} \tag{4}$$

[2]Different from the previous works in supervised pattern mining, we build the initial pattern candidates from images of the same object instance, rather than from the same target category. However, our pattern mining approach is general and also suitable for 2D patterns category-level supervision.

Transactions found for images $[I_1, \ldots, I_{n_t}]$ in the tth ToI with order k is defined as $\mathbf{T}^k(ToI_t) = \{\mathbf{T}^k(I_1), \ldots, \mathbf{T}^k(I_{n_t})\}$. The goal of pattern mining is to mine m_t patterns from $\{\mathbf{T}^k(ToI_t)\}_{k=1}^{K}$, i.e., $\mathbf{P}_t = \{P_1, \ldots, P_{n_t}\}$ from ToI_t. And in total we have $\{\mathbf{P}_t\}_{t=1}^{M}$, which are subsequently used to generate CBoP[3].

While the traditional pattern candidates are obtained by coding the 2D concurrences of visual words in individual reference images, we propose to search the k nearest neighbors in the 3D point cloud of each ToI. This point cloud is constructed by structure-from-motion over the reference images with bundle adjustment [18], which can be further used in other view-based 3D object analysis applications [59]. Figure 3 shows several exemplar 3D point clouds of landmarks as detailed in Section V.

Distance based Pattern Mining. Previous works in visual pattern mining mainly resort to **T**ransaction based **C**o-location pattern **M**ining (**TCM**). For instance, works in [10], [11], and [12] built transaction features by coding the k nearest words in 2D[4]. A transaction in TCM can be defined by coding the 2D spatial layouts of neighborhood words. Then frequent itemset mining algorithms like APriori [19] are deployed to discover meaningful word combinations as patterns, which typically check the pattern candidates from orders 1 to K.

TCM can be formulated as follows: Let $\{V_1, V_2, \ldots, V_m\}$ be the set of all potential items, each of which corresponds to a visual word in our case. Let $\mathbf{D} = \{\mathbf{T}_1, \mathbf{T}_2, \ldots, \mathbf{T}_n\}$ be the set of all transactions extracted as above, each is a combination of items in $\mathbf{V}$ after spatial coding. For simplification, we use $i \in [1, n]$ to denote all transactions discovered with orders 1 to K in ToI_1 to ToI_M. Let $\mathbf{A}$ be an "itemset" for a given transaction $\mathbf{T}$, we define the *Support* of an itemset as

$$support(\mathbf{A}) = \frac{\left|\{\mathbf{T} \in \mathbf{D} | \mathbf{A} \subseteq \mathbf{T}\}\right|}{|\mathbf{D}|}, \tag{5}$$

If and only if $support(\mathbf{A}) \geq s$, the itemset $\mathbf{A}$ is defined as a frequent itemset of $\mathbf{D}$, where s is the threshold to restrict the minimal support rate. Note that any two $\mathbf{T}_i$ and $\mathbf{T}_j$ are induplicated.

We then define the *confidence* of each frequent itemset as:

$$\begin{aligned} condifence(\mathbf{A} \rightarrow \mathbf{B}) &= \frac{support(\mathbf{A} \cup \mathbf{B})}{support(\mathbf{A})} \\ &= \frac{\left|\left\{\mathbf{T} \in \mathbf{D} | (\mathbf{A} \cup \mathbf{B}) \subseteq \mathbf{T}\right\}\right|}{\left|\{\mathbf{T} \in \mathbf{D} \middle| \mathbf{A} \subseteq \mathbf{T}\}\right|}, \end{aligned} \tag{6}$$

where $\mathbf{A}$ and $\mathbf{B}$ are two itemsets. The *confidence* in Equation 6 is defined as the maximal likelihood that $\mathbf{B}$ is correct in the case that $\mathbf{A}$ is also correct. The *confidence*-based restriction is to guarantee that the patterns can discover the minimal item subsets to represent the visual features at order $k \in [2, K]$.

[3]Note that the k nearest neighbor could be either 2D or 3D. And it will be refined later by a *gravity* distance metric.

[4]This spatial configuration can be further refined by incorporating the scales of interest points, which imposes scale invariance into transactions [13]

To yield a minimal association hyperplane to bound $\mathbf{A}$, an *Association Hyperedge* of each $\mathbf{A}$ is defined as:

$$AH(\mathbf{A}) = \frac{1}{N} confidence\big((\mathbf{A} - \{V_i\}) \rightarrow V_i\big). \tag{7}$$

Finally, by checking all possible itemset combinations in $\mathbf{D}$ from order 2 to K, the itemsets with $support() \geq s$ and $AH \geq \gamma$ are defined as frequent patterns.

One important issue of TCM relates to the repeated patterns in texture regions containing dense words. To address this, **D**istance-based **C**o-location Pattern **M**ining (**DCM**) is proposed with two new measures named *p*articipation *r*atio (*pr*) and *p*articipation *i*ndex (*pi*).

First, a *R-reachable* measure is introduced as the basis of both *pi* and *pr*: Two words V_i and V_j are *R-reachable* when

$$dis(V_i, V_j) < d_{thres}, \tag{8}$$

where $dis()$ is the distance metric in the spatial space, such as Euclidean and d_{thres} is the distance threshold. Subsequently, for a given word V_i, we define its partition rate $pr(\mathbf{V}, V_i)$ as the percentage of subset $\mathbf{V} - \{V_i\}$ that are R-reachable:

$$pr(\mathbf{V}, V_i) = \frac{\pi\left(\left|instance(\mathbf{V})\right|\right)}{\left|instance(V_i)\right|}, \tag{9}$$

where π is the relational projection operation (a function to operation on $instance(\mathbf{V})$) with de-duplication. The participation index *pi* is defined as:

$$pi(\mathbf{V}, V_i) = min_{i=1}^{m}\{pr(\mathbf{V}, V_i)\}, \tag{10}$$

where *pi* describes the frequency of subset $\mathbf{V} - V_i$ in the neighborhood. Note that only item subsets with *pi* larger than a give threshold is defined as patterns in **DCM**.

Gravity Distance R-reachable: In many cases, the Euclidean distance cannot discriminatively describe the co-location visual patterns, as the word discriminability and scale in item construction are ignored. Intuitively, words from the same scale tend to share more visual senses in item construction, and more discriminative words also produce more meaningful items. With this principle, we proposed a novel *Gravity Distance R-reachable* (*GD R-reachable*) to incorporate both cues. Two words V_i and V_j are GD R-reachable once $R_{i,j} < Cr_ir_j$ in the traditional **DCM** model, where r_i and r_j are the local feature scales of V_i and V_j respectively, C is the fixed parameter and $R_{i,j} = dis(V_i, V_j)$ is the Euclidean distance of two words.

To help interpretation, we can image that every word has a certain "gravity" to the other words, which is proportional to the square of its local feature scale. If this gravity is larger than a minimal threshold F_{min}, we denote these two words as *GD R-reachable*:

$$\begin{aligned} F^{i,j} &= \varepsilon \frac{\pi(r_i)^2 \pi(r_j)^2}{(R_{i,j})^2}, \varepsilon \; is \; a \; constant \\ F_{i,j} > F_{min} &\rightarrow \varepsilon \frac{\pi(r_i)^2 \pi(r_j)^2}{(R_{i,j})^2} > F_{min} \rightarrow R_{i,j} < Cr_ir_j. \end{aligned} \tag{11}$$

Similar to **DCM**, the input of gravity-distance-based mining is all instances of visual words. Each instance contains the

Algorithm 1: Gravity distance based visual pattern mining

Input: Visual vocabulary $\mathbf{V}$, reference images $\{I_i\}_{i=1}^{N}$ reference images with respect to ToIs $\{\{I_i^{ToI_t}\}_{i=1}^{n_t}\}_{t=1}^{M}$, bag-of-words histograms $\{\mathbf{V}(1), ...\mathbf{V}(N)\}$, $support$ threshold s, $confidence$ threshold γ, maximal pattern order K, and sparse factor α.
Output: CBoP pattern set $\{Q\}_{i=1}^{n_{selected}}$.
// **3D Sphere Coding**:
for the tth ToI ($t \in [1, M]$) **do**
  Build 3D point cloud using structure-from-motion with bundle adjustment;
  Build transactions $\{\mathbf{D}\}_t$ by 3D sphere coding in the point cloud with *GD R-reachable distance*;
end
Ensemble $\{\mathbf{D}\}_t$ for all ToI ($t \in [1, M]$) as $\mathbf{D} = \{\mathbf{D}\}_{t=1}^{M}$;
// **Gravity Distance based Pattern Mining**:
Calculate itemset supports and confidences by Equation 5, 6;
Filtering out unreliable pattern candidates by s and γ;
Ouput patterns $\mathbf{P} = \{\mathbf{P}_t\}_{t=1}^{M}$ by Equation 10, 9;
// **Sparse Pattern Coding**:
Learn the coding vector $\mathbf{w}$ from $\mathbf{P}$ by Equation 12, 15.
if $\{L_i\}_{i=1}^{N}$ are also available **then**
  Conduct supervised coding by Equation 14.
end

following attributes: *original local features*, *visual word ID*, *location* and *scale* of this local feature (word instance). To unify the description, we embed the *word ID* of each feature with its corresponding *location* into the mining. Then, we run the same procedure of **DCM** to mine co-location patterns. Algorithm 1 outlines the work flow of our gravity distance based visual pattern mining. Figure 4 shows some studied cases of the mined patterns.

C. Sparse Pattern Coding

Sparse Coding Formulation. Given the mined pattern collection $\mathbf{P}$, not all patterns are equivalently important and discriminative. Indeed, there are typically redundancy and noise in this initial pattern set. However, how to design a compact yet discriminative pattern-level features are left open in the literature. In this subsection, we formulate the pattern-level representation as a sparse pattern coding problem, aiming to maximally reconstruct the original bag-of-words histogram using a minimal number of patterns.

Formally speaking, let $\mathbf{P} = \{P_1, \ldots, P_L\}$ be the mined patterns with maximal order K. In online coding, given the bag-of-words histogram $\mathbf{V}(q) = [V_1(q), \ldots, V_m(q)]$ extracted from image I_q, we aim to encode $\mathbf{V}(q)$ by using a compact yet discriminative subset of patterns, say $\mathbf{P}(q) \subset \mathbf{P}$. This is formulated as seeking an optimal tradeoff between maximizing the reconstruction ability and minimizing the coding length:

$$\arg\min_{\mathbf{w}} \sum_{i=1}^{N} ||\mathbf{V}(i) - \mathbf{w}^T \mathbf{P}(i)||_2 + \alpha||\mathbf{w}||_1, \tag{12}$$

where $\mathbf{P}$ is the patterns mined previously, from which a minimal set is selected to lossy reconstruct each bag-of-words histogram $\mathbf{V}(i)$ as much as possible. $\mathbf{w}$ serves as a weighted linear combination of all non-zero patterns in $\mathbf{P}$, which pools patterns to encode each $\mathbf{V}(i)$ as:

$$f_{\mathbf{P}}(i) = w_1 P_1 + w_2 P_2 +, \ldots, w_m P_m, \tag{13}$$

where $[w_1, w_m]$ is the learnt weighted vector to reconstruct $\mathbf{V}(i)$ in Equation 12. Each w_j is assigned to either 0 or 1, performing a binary pattern selection. The summed up is operated as follows: First, we expand each transaction to a full-length BoW, in which only words corresponding to the itemset is non-zero. Then, these "expanded" patterns are summed up, producing the final overall BoW.

Learning with respect to Equation 12 and 15 are achieved by spare coding over the pattern pool $\mathbf{P}$. While guaranteeing the real sparsity through $\mathcal{L}_0$ is intractable, we approximate a sparse solution for the coefficients $\mathbf{w}$ using $\mathcal{L}_1$ penalty, which results in a Lasso based solution [45].

Finally, we denote the selected pattern subset as $\mathbf{Q}_{selected}$, which contains $n_{selected}$ patterns as $[Q_1, \ldots, Q_{n_{selected}}]$. $n_{selected}$ is typically very small, say one hundred. Therefore, each reference or query image is represented as an $n_{selected}$-bin pattern histogram.

Supervised Coding. Once the supervision labels $\{L_i\}_{i=1}^{N}$ (e.g. category label or prediction) for reference image $\{I_i\}_{i=1}^{N}$ are also available, we can further incorporate L_i to refine the coding of Equation 12 as:

$$\arg\min_{\mathbf{w}} \sum_{i=1}^{N} ||\mathbf{V}(i) - (\mathbf{w}^T \mathbf{u})^T \mathbf{P}(i)||_2 + \alpha||\mathbf{w}||_1, \tag{14}$$

where we integrate the supervised label $\{L_i\}_{i=1}^{N}$ into the coding function of $\mathbf{P}$ for $I(i)$ as:

$$(\mathbf{w}^T \mathbf{u})^T \mathbf{P}(i) = w_1 u_1 P_1 + w_2 u_2 P_2 +, ..., w_m u_m P_m. \tag{15}$$

$[u_1, \ldots, u_m]$ adds the prior distribution of patterns to weight the selection of $[w_1, \ldots, w_m]$, where u_j is the discriminability of the jth pattern based on its information gain to L_i:

$$u_j = H(L_i) - H(L_i|P_j) \tag{16}$$

Here, $H(L_i)$ is the prior of label L_i given I_i, $H(L_i|P_j)$ is the conditional entropy of label L_i given P_j, which is obtained by averaging the intra-class observation of P_j divided by the inter-class distribution of P_j:

$$H(L_i|P_j) = \frac{p(P_j|L_j)}{\sum_l p(P_j|L_l)}. \tag{17}$$

In this sense, the definition of supervised label L_i is fairly flexible, e.g., category labels or positive/negative tags, which also allows the case of missing labels.

In terms of compact visual descriptor, CBoP preserves the higher-order statistics comparing to the transitional non-zero coding scheme [20] as well as the state-of-the-art boosting-based word selection scheme [17]. Different from all previous unsupervised descriptor learning schemes [10]–[13], and [20], CBoP further provides a supervised coding alternative as an

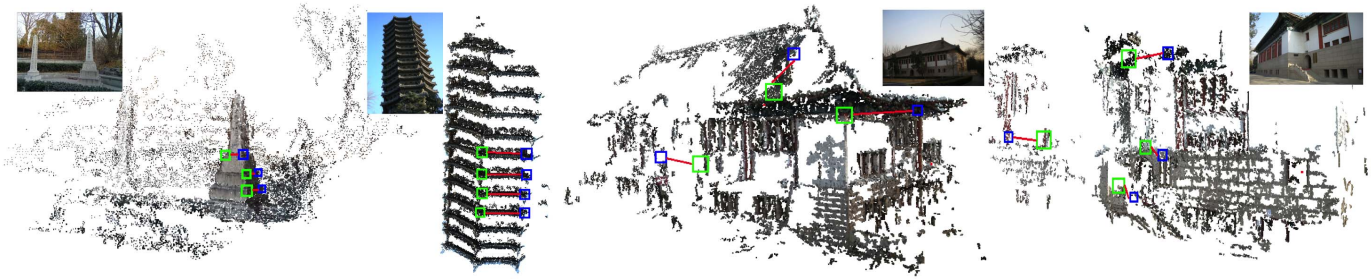

Fig. 4. Case study of the mined patterns between the gravity distance based pattern mining and the Euclidean distance based pattern mining.

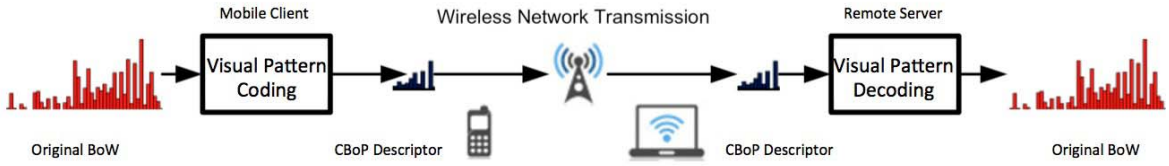

Fig. 5. The proposed low bit rate mobile visual search framework using CBoP descriptor. Different from previous works in near-duplicate visual search, we emphasize on extremely compact descriptor extraction directly on the mobile end. Our CBoP descriptor is typically of hundreds of bits. To the best of our knowledge, the CBoP is the most compact descriptor with comparable discriminability to the state-of-the-art visual descriptors [2], [20], [21], [26].

optional choice, which yet differs from the work in [17] that demands online side information from the query.

IV. CBoP for Low Bit Rate Mobile Visual Search

We demonstrate the advantages of the proposed CBoP descriptor in the scenario of low bit rate mobile landmark search. Different from sending the query image compact descriptors are directly extracted and sent from the mobile end to reduce query delivery latency. Such descriptor(s) is expected to be compact, discriminative, and can be efficiently extracted.

Search Pipeline: Figure 5 shows the work flow of using our CBoP descriptor in this prototype. The algorithm extracts local features from the query image, quantizes them into bag-of-words, scans 2D spatial nearby words to generate potential patterns, and encodes the discovered patterns into CBoP [5]. We further compress this descriptor into a bin occurrence (hit/non-hit) histogram with residual coding.

In the remote server, the decoding procedure is performed: First, a difference decoding is conducted to obtain CBoP histogram, which is then recovered into a bag-of-words histogram by summing up all non-zero patterns using their word combinations with weights:

$$\mathbf{V}(q)_{recovered} = \mathbf{w}^T \mathbf{Q} = \sum_{i=1}^{n_{selected}} w_i Q_i \tag{18}$$

The decompressed BoW is subsequently sent to the VT-based search system where the ranking is conducted. Note that the spatial layouts of words can be also sent to conduct re-ranking.

Efficiency. In our current implementation, given a query image, we only need approximately 2 seconds to extract the CBoP descriptor. By using visual word pruning technique (e.g. only maintains the visual word centroid features as well as an approximate radius of each word), the storage cost of VT is also limited.

Contextual Learning. There is cheaply available side information in the mobile end, such as GPS tags, compass direction and base station identity tag. Exploiting the above side information with our supervised coding (Section III-C), the extracted CBoP can be further refined, as detailed in the subsequent experiments (Section V).

V. Quantitative Results

A. Data Collection

PhotoTourism. First, we perform the patch level validation over the the image patch correspondence benchmark[6]. It contains over 100,000 image patches with correspondence labels generated from the point clouds of *Trevi Fountain* (Rome), *Notre Dame* (Paris) and *Half Dome* (Yosemite), all of which are produced by the Photo Tourism system [18]. Each patch correspondence consists of a set of local patches, which is obtained by projecting a given 3D point from the point cloud back to multiple reference images and cropping the corresponding salienct regions[7]. Some exemplar patches obtained through the above procedure are shown in Figure 6.

10M Landmark Photos. To validate our CBoP descriptor in the scalable image retrieval application, we have also collected over 10 million geo-tagged photos from both Flickr (http://www.Flickr.com) and Panoramio (http://www.Panoramio.com) websites. We crawled photos tagged within five cities *i.e. Beijing*, *New York City*, *Lhasa*, *Singapore* and *Florence*. This dataset is named as *10M Landmarks*. Some exemplar photos are shown in Figure 7.

We use k-means clustering to partition photos of each city into multiple regions based on their geographical tags. For

[5]Note that while the patterns are built offline through 3D sphere coding, in online search we use their 2D codes since we do not have a structure correspondence from the query image to the reference point clouds. Not doubt, this introduces certain distortion. However, as shown in our experiments, superior performance over state-of-the-arts can be still guaranteed.

[6]http://phototour.cs.washington.edu/datasets/

[7]Since this dataset contains the ground truth patch correspondences as well as the point clouds from [18] and [48], we will skip the the 3D sphere coding (Section III) in the following quantitative tests.

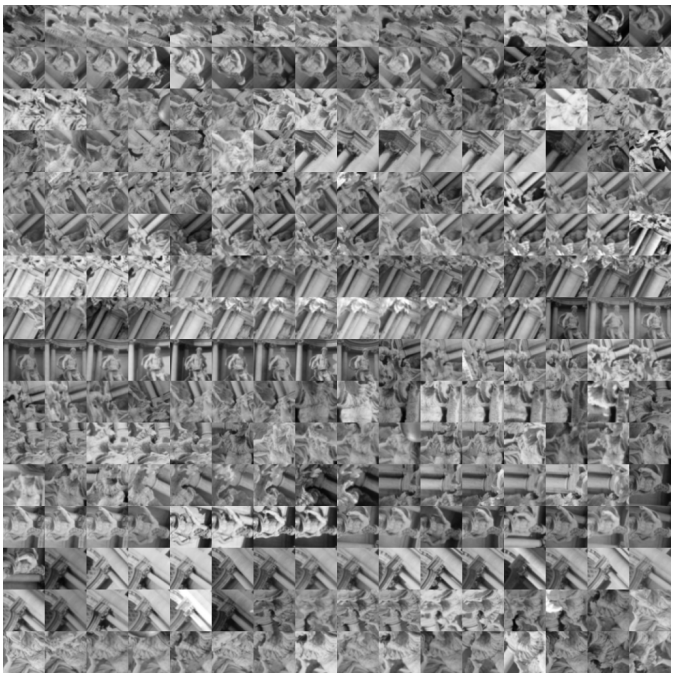

Fig. 6. Exemplar local patches in the PhotoTourism dataset. Each patch is sampled at patch size 64×64 with a canonical scale and orientation. For details on how the scale and orientation is established, please refer to [48]. These ground truth correspondences are collected from the structure-from-motion based point cloud construction, through the back projection of the matched points.

each city, we select the top 30 densest regions as well as 30 random regions. We then invite a group of volunteers to identify one or more dominant landmark views for each of these 60 regions. For an identified dominant view, all their near-duplicate photos are manually labeled in its belonged region and nearby regions. Eventually, we have 300 queries as well as their ground truth labels (corrected matched photos) as our test set.

PKUBench. We also extend the validation to the typical mobile search scenario over *PKUBench* (http://61.148.212.146:8080/landmark/benchmark), which is a public available mobile visual search benchmark. This benchmark contains 13,179 scene photos organized into 198 landmark locations, captured by both digital and phone cameras. There are in total 6,193 photos captured from digital cameras and 6,986 from phone cameras respectively. We recruited 20 volunteers in data acquisition. Each landmark is captured by a pair of volunteers with a portable GPS device (one using digital camera and the other using phone camera). Especially, this benchmark includes four groups of exemplar mobile query scenarios (in total 118 images):

- *Occlusive Query Set*: 20 mobile queries and 20 corresponding digital camera queries, occluded by foreground cars, people and buildings.
- *Background Cluttered Query Set*: 20 mobile queries and 20 corresponding digital camera queries, captured far away from a landmark, where GPS search yields worse results due to the signal bias of nearby buildings.
- *Night Query Set*: 9 mobile phone queries and 9 digital camera queries, where the photo quality heavily depends on the lighting conditions.
- *Blurring and Shaking Query Set*: 20 mobile queries with blurring or shaking and 20 corresponding digital camera queries without any blurring or shaking.

B. Evaluation Criteria

Effectiveness. We use mean Average Precision (*mAP*) to evaluate our performance, which is also widely used in the state-of-the-art works [1], [3], [4], [21], [26]. *mAP* reveals the position-sensitive ranking precision by the returning list:

$$MAP@N = \frac{1}{N_q}\sum_{i=1}^{N_q}\left(\frac{\sum_{r=1}^{N} P(r)rel(r)}{min(N, \#-relevant-images)}\right) \tag{19}$$

where N_q is the number of queries; r is the rank, N the number of related images for query i; $rel(r)$ a binary function on the relevance of r; and $P(r)$ the precision at the cut-off rank of r.

Note that here we have a min operation between the top N returning and $\#-relevant-images$. In large-scale search, there are always over hundreds of ground truth relevant images to each query. Therefore, dividing by $\#-relevant-images$ would result in a very small MAP. Alternatively, a better choice is to divide by the number of returning images. We use $min(N, \#-relevant-images)$ to calculate MAP@N[8].

Efficiency. We evaluate the efficiency using the descriptor compactness, i.e., the size of the descriptor stored in the memory or hard disk (for instance 1KB per descriptor etc.).

Extreme Cases. We show our performance in several extreme mobile visual search scenarios, i.e., blurred queries or occlusive queries, as detailed in Section V-G.

C. Baselines

- *Bag-of-Words*: Transmitting the entire BoW incurs the lowest compression rate. However, it provides an mAP upper bound to the other BoW compression strategies. Note that in practice considering only non-zero words are transmitted, more practical solutions such as Tree Histogram Coding (detailed later) would be used.
- *2D Patterns*: Instead of obtaining the initial pattern candidates through 3D sphere coding, using the point cloud, we adopt the traditional 2D spatial coding from individual reference images as initially proposed in [13]. This baseline validates the effectiveness of our proposed 3D sphere coding strategy.
- *LDA [23]*: One straightforward alternative in abstracting compact features from the bag-of-words histogram is the topic model features [7], [22], [23]. In this paper, we implement the well-used Latent Direchellet Allocation (LDA) based feature in [23] as our baseline.
- *CHoG [21]*: We also compare our CBoP descriptor to the CHoG local descriptor, which is a state-of-the-art alternative as introduced in [21].

[8]In the case that N is smaller than the number of labeled ground truth, we can simplify $min(N, \#-relevant-images)$ with N in the subsequent calculation. The $min(N, \#-relevant-images)$ operation is a common evaluation in the TRECVID evaluation (http://trecvid.nist.gov/).

Fig. 7. Exemplar photos collected from Flickr and Panoramio to construct our *10M landmarks* dataset.

- *Tree Histogram Coding [20]*: In terms of image-level compact descriptor, we further implement the Tree Histogram Coding scheme proposed by Chen et al. [20], which is a lossless BoW compression that uses residual coding to compress the BoW histogram[9].
- *LDVC [17]*: Finally, we compare our CBoP descriptor with the state-of-the-art compact descriptor to the best of our knowledge [17], which adopts location discriminative coding to boost a very compact subset of words. To compare to our unsupervised CBoP, we used reference images from the entire dataset (i.e. landmark photos from the entire city) to train the LDVC descriptor. To compare to our supervised CBoP, we adopt the location label to train a location sensitive descriptor as in [17].

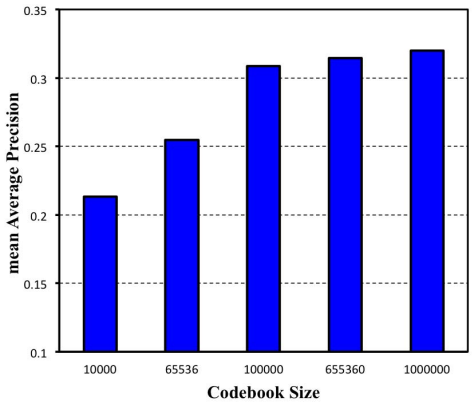

Fig. 8. The influence of codebook size on the CBoP based mAP performance.

D. Implementation Details

Settings. For each reference dataset, there are four steps to build our CBoP offline: *Step I*: Vocabulary Construction, *Step II*: 3D Sphere Coding, *Step III*: Visual Pattern Mining, and *Step IV*: CBoP construction. For both *10M Landmarks* and *PKUBench*, we run bundle-based structure-from-motion to obtain the patch correspondence set and transaction configurations and do pattern mining and CBoP construction, which covers *Step I* to *Step IV*. For *PhotoTourism*, it already provides the patch correspondence set as well as the point clouds, which can be directly treated as the input of our CBoP descriptor learning. Therefore, only *Step IV* and the mining part of *Step III* are involved.

Million Scale Vocabulary. We adopt dense sampling to extract local features and described by SIFT [37]. We then adopt the Vocabulary Tree model [2] to build our initial visual vocabulary with inverted indexing[10]. Figure 8 shows how different codebook sizes affect the pattern mining results. In conclusion, we set the branching factor $b = 10$ and the tree depth $h = 6$ in our final implementation, which produces approximate $0.1M$ words.

Visual Pattern Mining. *Structure from Motion*: Given the reference images captured at multiple viewing angles for each ToI, we adopt structure-from-motion to simultaneously estimate the 3D point matching and the intrinsic camera parameters of these images. Figure 9 shows several examples of 3D point clouds estimated by structure-from-motion. Similar to [18], we extract interest points and match them between images using SIFT [37]. Then, candidate matches are done using a symmetry criterion and the initial fundamental matrices between image pairs are estimated by RANSAC plus bundle adjustment [18].

Word Quantization. We project each virtual 3D point back into individual images, if and only if this point was randomly sampled at the very beginning and retained after structure-from-motion estimation. We sample 64×64 pixels around each interest point as input patches, which are quantized into visual words. Finally, for each 3D virtual point, the majority of the visual word bin (there could be multiple bins since these patches could be quantized into different words) from its reference image patches is selected to represent the word identity of this 3D point.

3D Sphere Coding. Subsequently, for each 3D virtual point, we scan its K spatial neighborhood in 3D to build the initial itemset collection.

Visual Pattern Mining based on GD R-reachable DCM. We adopt the DCM based pattern mining scheme with GD R-reachable distance (Section III) to mine visual patterns from each ToI. The ensemble of all mined patterns forms the pattern pool for the subsequent sparse pattern coding (The quantitative comparisons between the Euclidean distance and Gravity distance are detailed later.).

Sparse Pattern Coding Implementations. The sparsity factor α in Section III-C acts as a tradeoff between the recon-

[9]Please note that, THC is a lossless compression, which can be decoded as using the original visual words as query. In THC, this is done by only recording the non-zero word frequencies with their positive difference (using difference coding). We compare to THC to see the mAP variations of using the original visual words from a query image vs. using the set of visual words decoded from the visual patterns.

[10]We adopt the VT model to build the initial bag-of-words histograms, but our solution is general enough for other vocabularies such as k-means clustering [1], Approximate K-Means [3] and their variances [25], [3] etc.

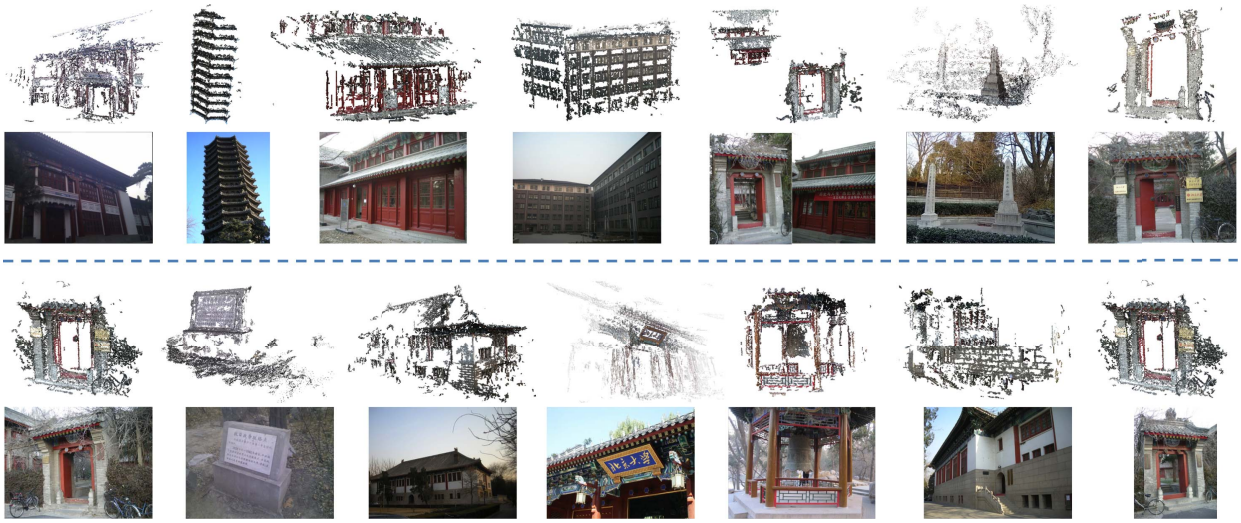

Fig. 9. Examples of the reconstructed point clouds from touristic landmarks in Peking University.

struction capability of the original bag-of-words histogram and the compactness of the CBoP descriptor. For simplicity, we directly adopt a ten-fold cross validation on our training set to figure out the best α for both offline and online coding.

E. Insights Into CBoP

Pattern Discriminability. We investigate the rationality of CBoP by running boosting based feature selection over both visual patterns $\mathbf{P}$ and visual words $\mathbf{V}$. Ideally, the mined patterns should be more discriminative than the original words when measuring them together. Our evaluation procedure is similar to the feature selection approach as described in [17]: First, we sample a set of training images ($[\mathbf{I}'_1, \ldots, \mathbf{I}'_{n_{sample}}]$) and use each image to search similar images by its bag-of-words histogram, which results in the following training set:

$$\begin{aligned} Query(\mathbf{I}'_1) &= [\mathbf{A}_1^1, \mathbf{A}_2^1, \ldots, \mathbf{A}_R^1] \\ \ldots &= \ldots \\ Query(\mathbf{I}'_{n_{sample}}) &= [\mathbf{A}_1^{n_{sample}}, \mathbf{A}_2^{n_{sample}}, \ldots, \mathbf{A}_R^{n_{sample}}] \end{aligned} \tag{20}$$

where $\mathbf{A}_i^j$ is the ith returning of the jth query. We expect the boosted words/patterns from $\mathbf{P} \cup \mathbf{V}$ should maximally retain $[\mathbf{A}_1^j, \mathbf{A}_2^j, \ldots, \mathbf{A}_R^j]$ for each jth query.

To this end, we define $[w_1, \ldots, w_{n_{sample}}]$ as an error weighting vector to the n_{sample} images used as pesudo query, which measures the ranking consistency loss in the word/pattern selection. We further define the selected word/pattern subset as $\mathbf{C}$. At the tth boosting iteration, we got the current $(t-1)$ words/patterns as $\mathbf{C}^{t-1}$. To select the next tth discriminative word or pattern from $\mathbf{P} \cup \mathbf{V}$, we estimate the ranking preservation of the current selection $\mathbf{C}^{t-1}$ as:

$$Loss(\mathbf{I}'_i) = w_i^{t-1} \sum_{r=1}^{R} Rank(\mathbf{A}_r^1) \mathbf{W}_{\mathbf{A}_r^1} || f(\mathbf{C}_{\mathbf{I}'_i}^{t-1}), \mathbf{V}_{\mathbf{A}_r^i} ||_2 \tag{21}$$

where $i \in [1, n_{sample}]$; $Rank(\mathbf{A}_r^i)$ is the current position of the originally ith returning of query $\mathbf{I}'_i$; $f()$ the bag-of-words recovery function similar to Equation 18. $[w_1^{t-1}, \ldots, w_{n_{sample}}^{t-1}]$ the *(t-1)*th error weighting, which measures the ranking loss

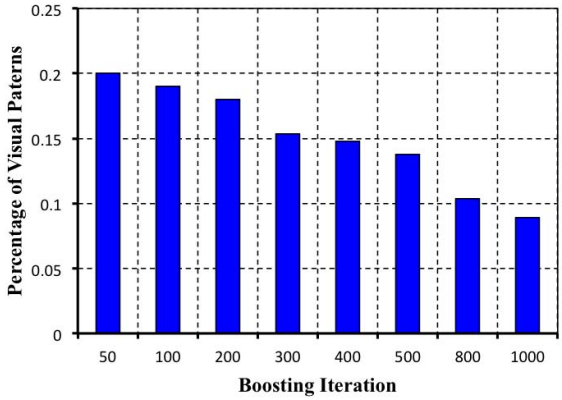

Fig. 10. Percentage of visual patterns left in the feature pool (visual words + visual patterns) after the Boosting based selection. Here, a lower percentage means a higher percentage of patterns being selected into the final CBoP descriptor.

of the jth query ($j \in [1, n_{sample}]$). Then, the overall ranking loss is:

$$Loss_{Rank} = \sum_{i=1}^{n_{sample}} w_i^{t-1} \sum_{r=1}^{R} Rank(\mathbf{A}_r^i) \mathbf{W}_{\mathbf{A}_r^i} || f(\mathbf{C}_{\mathbf{I}'_i}), \mathbf{V}_{\mathbf{A}_r^i} ||_2 \tag{22}$$

The next word/pattern C_t is selected based on:

$$\begin{aligned} C_t = \arg\min_j & \sum_{i=1}^{n_{sample}} w_i^{t-1} \sum_{r=1}^{R} Rank(\mathbf{A}_r^i) \mathbf{W}_{\mathbf{A}_r^i} \\ & \times || f(\{\mathbf{C} + C_j\}_{\mathbf{I}'_i}), \mathbf{V}_{\mathbf{A}_r^i} ||_2 \end{aligned} \tag{23}$$

Subsequently, we update the error weighting of each w_i^{t-1} as the corresponding loss of each ith query in Equation 23.

Figure 10 shows the percentage of visual patterns learnt by using our Boosting based selection. A pattern level feature is an important proposition in the boosted feature.

Pattern Order. Another issue is to determine the max pattern order K, which controls the tradeoff between the time efficiency and the CBoP effectiveness. To a certain degree, it also controls the tradeoff between the descriptor generality

TABLE I
MEMORY (MB) REQUIREMENTS FOR CBoP AND OTHER ALTERNATIVES ON THE MOBILE END

Tree Structure	Memory Requirement
SIFT SVT, H = 6, B = 10	59MB
CHoG SVT, H = 6, B = 10	24MB
Aggregate Local Descriptors [26]	728MB
CBoP, , H = 6, B = 10	59MB

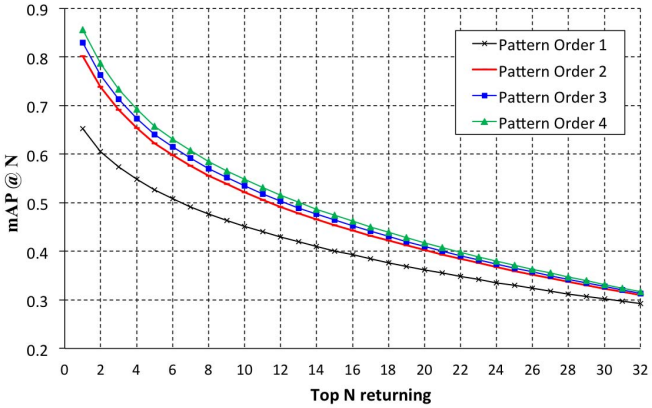

Fig. 11. *mAP* variances with respect to the maximal pattern order K. For each K, a fixed CBoP length 1,000 is applied to yield the best performance.

and the descriptor discriminability. For instance, higher-order patterns would lead to the over fitted descriptor learning over the training set[11]. By increasing the maximal pattern order, the computation cost would be higher. Figure 11 shows how the mAP varies with respect to different maximal pattern order K. It is clear that from the results of pattern orders 3 and 4, the order influence is limited.

Note that VT compression is applied: First, we applied [52] to do VT pruning, in which most of less useful subtrees are pruned out in the online quantization. Second, we have observed that the central SIFT points of nearby words in VT are typically repeatable in many dimensions. Therefore, difference coding is applied to further compress word centers at each layer.

F. Quantitative Performance

Efficiency Analysis. We deploy the low bit rate mobile visual search prototype on HTC Desire G7 as a software application. The HTC DESIRE G7 is equipped with an embedded camera with maximal 2592 × 1944 resolution, a Qualcomm MSM7201A processor at 528MHz, a 512M ROM + 576M RAM memory, 8G extended storage and an embedded GPS. Tables I and II show the memory and time cost with comparisons to state-of-the-arts in [20], [21], and [26]. In our CBoP descriptor extraction, the most time-consuming part is the local feature extraction, which can be further accelerated by random sampling, instead of using the interest point detectors [21], [37].

[11]In an extreme case, if there is only one training image, we can simply retain all non-zero words into one pattern, producing a 1-bit descriptor for this ToI.

TABLE II
TIME (SECOND) COST FOR CBoP AND OTHER ALTERNATIVES ON THE MOBILE END

Compression Methods	Local Feature	Descriptor Coding
BoW Histogram	1.25S	0.14S
Aggregate Local Descriptors [26]	1.25S	164S
Tree Histogram Coding [20]	1.25S	0.14S
Vocabulary Boosting[17]	1.25S	0.14S
CBoP	1.25S	0.14S

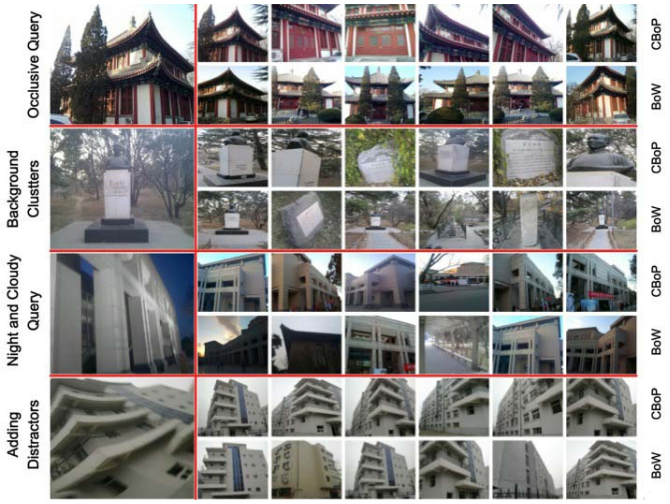

Fig. 12. Query example comparisons in extreme mobile query scenarios including *Occlusive Query Set*, *Background Cluttered Query Set*, *Night Query Set* and *Blurring and Shaking Query Set* in the *PKUBench* dataset.

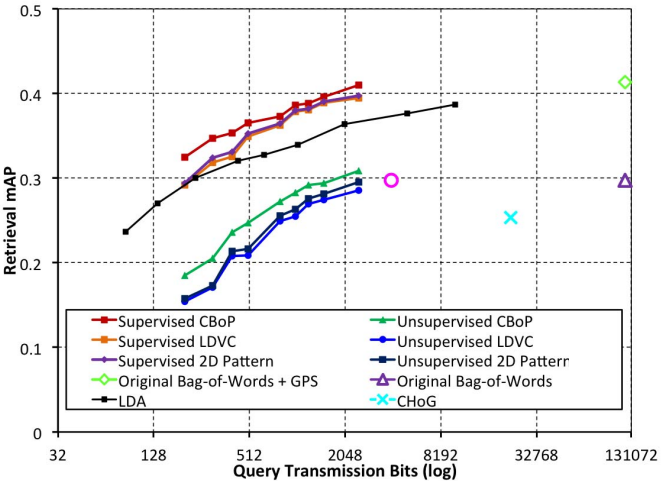

Fig. 13. Distortion analysis of retrieval performance versus query compression rates, as well as the comparisons to [20], [21], and [26] using the ground truth query set (the pink circle corresponds to the THC [20]).

Rate Distortion Analysis. To compare our CBoP descriptors to the baselines [20], [21], [26], we perform the rate distortion analysis in Figure 13, where the rate means the descriptor lengths of our CBoP descriptor and other alternatives, while the distortion is measured by the search performance drop in terms of mAP with respect to different methods. As shown in Figure 13, our CBoP descriptor has achieved the best tradeoff in the rate distortion evaluation. It reports the highest compression rate with a comparable distortion (by viewing Figure 13

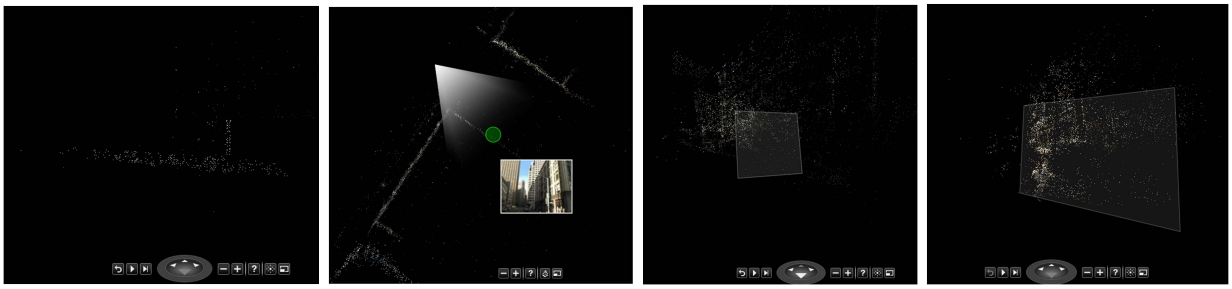

Fig. 14. Visualized examples of sparse and dense point clouds in the San Francisco Landmark Dataset [58]. From left to right: Sparse point cloud, overhead view of point clouds on the street, dense point cloud example 1, dense point cloud example 2.

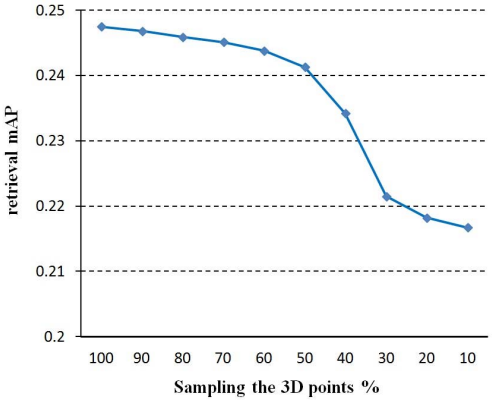

Fig. 15. The correlation between the percentages of sampled 3D points and the retrieval mAP using CBoP.

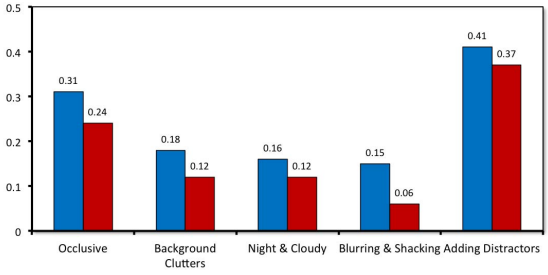

Fig. 16. mAP comparisons of our CBoP descriptor (blue bar) to its original BoW features (red bar) in the extreme mobile query scenarios in *PKUBench*.

horizontally), as well as the highest ranking performance with a comparable compression rate (by viewing Figure 13 vertically). In addition, without supervised learning, our CBoP descriptor can still achieve better performance comparing to all alternatives and state-of-the-arts [20], [21], [26].

Note that the baselines of 2D Patterns, LDA, and THC are all based on 2D images rather than 3D point clouds. Therefore, the superior performance over these baselines has demonstrated the effectiveness of mining patterns from 3D point clouds.

Extreme Mobile Query Scenarios. Figure 12 further shows five cases of imperfect mobile visual query scenarios in *PKUBench*, including *Occlusive Query Set*, *Background Cluttered Query Set*, *Night Query Set* and *Blurring and Shaking Query Set*. Our CBoP descriptor achieves identical or better performance, due to its higher-order statistics to capture the Eigen structure of scene/object against photographing changes. Figure 16 shows the performances of our CBoP and other alternatives in the extreme mobile query scenarios in the *PKUBench*. As in Figure 16, it is obvious that the mined patterns are much more discriminative even comparing to the BoW features, in terms of its discriminability between foreground and background words, as well as the fact that most patterns are located in the most discriminative regions from the target of interest.

On the Application Scenario of CBoP: Although our approach is designed for large-scale mobile landmark search, not every landmark dataset works well as our case. Especially, given our 3D sphere coding from the 3D point clouds, the dataset itself has to involve close and related viewing angles for every landmark. This is the case when preparing our landmark dataset (the 10M Landmark Dataset), in which the reference images of each landmark are captured from different but overlapped viewing angles. However, it is not always true for some other datasets such as the San Francisco dataset [58], which depends on the LadyBug scanning routes. For instance, buildings located at the cross-street should have more overlapped viewing angles hence the point clouds are dense (and therefore more visual patterns could be discovered), while building on the street might be not. As shown in Figure 14:

On the Sparsity of Point Clouds: Another practical issue would be influence of point cloud sparsity, which intuitively would affect the search performance of using CBoP descriptor. As shown in Figure 15, we progressively reduce (subsample) the number of 3D points in building upon the point clouds and evaluate the retrieval mAP degeneration. It is obvious that there is a significant drop when retaining about less than 30% points, while the performance remains stable with $>50\%$ points.

G. Case Study and Visualized Examples

Visualized Search Results with Patterns. Figure 17 further shows the performance of using CBoP histogram for the mobile visual search in our *10M Landmarks* dataset, with comparisons to both bag-of-words based representation as well as 2D pattern features. In Figure 17, our CBoP achieves very comparable search accuracy with a very limited number of histogram bins, e.g. hundreds, in contrast to the million-scale BoW histogram.

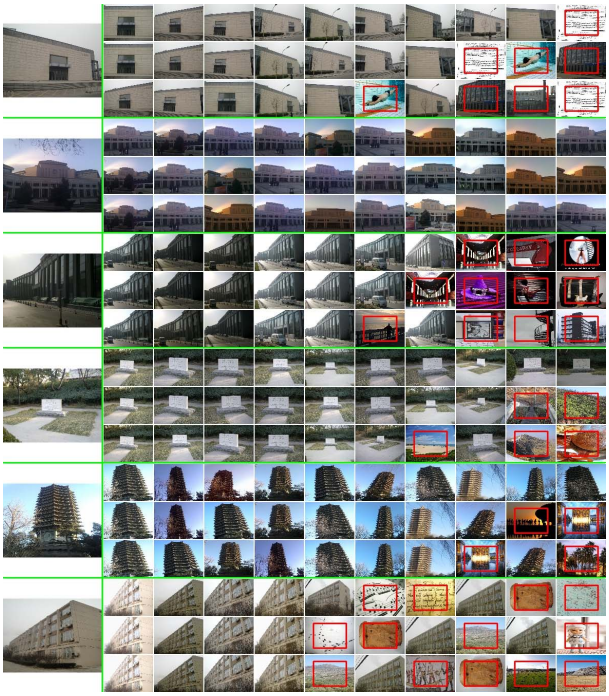

Fig. 17. Case study of retrieval in the *10M Landmarks*. The first image in each row is the query, and the right hand side shows retrieval results of: 1. CBoP, 2. 2-dimensional CBoP, and 3. BoW. Note that the red rectangle indicates an incorrect retrieval returning.

VI. Conclusion

We have proposed to mine discriminative visual patterns from 3D point clouds, based on which we learn a Compact Bag-of-Patterns (CBoP) descriptor. The 3D point clouds based mining schemes alleviates the ill-posed pattern configurations from the 2D photographic statistics of individual images. Beyond existing pattern level representation, the proposed CBoP offers a compact yet discriminative visual representation, which significantly contributes to low bit rate mobile visual search.

To discover more precise pattern configurations in the real-world, we propose to reconstruct 3D point clouds of search objects by using structure-from-motion based on bundle adjustment, in which a 3D sphere coding is applied to precisely capture the co-location statistics of words in 3D point clouds. A gravity-based distance is introduced to mine co-location patterns, which incorporates the spatial distances of visual words to derive the patterns containing more discriminative words. Based upon the mined patterns, we further propose to build a compact yet discriminative image representation at the level of meaningful patterns, named Compact Bag-of-Patterns (CBoP). CBoP adopts a sparse coding to pursue a maximal reconstruction of the original bag-of-words histograms with a minimal pattern coding length. Finally, labels can be further incorporated to improve the discriminability of CBoP descriptor in a supervised manner.

We have validated the proposed CBoP descriptor in a low bit rate mobile landmark search prototype. We quantitatively demonstrate the advantages of CBoP on both benchmark datasets and a 10-million landmark photo collection. Our CBoP descriptor has outperformed the state-of-the-art pattern mining schemes [13], topic features [23], and compact descriptors [17], [20], [21].

References

[1] J. Sivic and A. Zisserman, "Video google: A text retrieval approach to object matching in videos," in *Proc. 9th IEEE ICCV*, Oct. 2003, pp. 1470–1477.

[2] D. Nistér and H. Stewenius, "Scalable recognition with a vocabulary tree," in *Proc. IEEE Comput. Soc. Conf. CVPR*, Jun. 2006, pp. 2161–2168.

[3] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Object retrieval with large vocabulary and fast spatial matching," in *Proc. IEEE Conf. CVPR*, Jun. 2007, pp. 1–8.

[4] G. Schindler, R. Szeliski, and M. Brown, "City-scale location recognition," in *Proc. IEEE Conf. CVPR*, Jun. 2007, pp. 1–7.

[5] G. Salton and C. Buckley, "Term-weighting approaches in text retrieval," *Inform. Process. Manag.*, vol. 24, no. 5, pp. 513–523, 1988.

[6] T. Hofmann, "Unsupervised learning by probabilistic latent semantic analysis," *J. Mach. Learn.*, vol. 42, nos. 1–2, pp. 177–196, Jan./Feb. 2001.

[7] D. Blei, A. Ng, and M. Jordan, "Latent dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Jan. 2003.

[8] Z. Wu, Q. Ke, M. Isard, and J. Sun, "Bundling features for large scale partial-duplicate web image search," in *Proc. IEEE Conf. CVPR*, Jun. 2009, pp. 25–32.

[9] J. Yang, K. Yu, Y. Gong, and T. Huang, "Linear spatial pyramid matching using sparse coding for image classification," in *Proc. IEEE Conf. CVPR*, Jun. 2009, pp. 1794–1801.

[10] J. Sivic and A. Zisserman, "Video data mining using configurations of viewpoint invariant regions," in *Proc. IEEE Comput. Soc. Conf. CVPR*, Jul. 2004, pp. I-488–I-495.

[11] T. Quack, V. Ferrari, and L. V. Gool, "Video mining with frequent itemset configurations," in *Proc. 5th Int. Conf. CIVR*, Jul. 2006, pp. 360–369.

[12] J. Yuan, Y. Wu, and M. Yang, "Discovery of collocation patterns: From visual words to visual phrases," in *Proc. IEEE Conf. CVPR*, Jun. 2007, pp. 1–8.

[13] T. Quack, V. Ferrari, B. Leibe, and L. V. Gool, "Efficient mining of frequent and distinctive feature configurations," in *Proc. IEEE 11th ICCV*, Oct. 2007, pp. 1–8.

[14] Y. Huang, S. Shekhar, and H. Xiong, "Discovering collocation patterns from spatial data sets: A general approach," *IEEE Trans. Knowl. Data Eng.*, vol. 16, no. 12, pp. 1472–1485, Dec. 2004.

[15] R. Fergus, P. Perona, and A. Zisserman, "Object class recognition by unsupervised scale-invariant learning," in *Proc. IEEE Comput. Soc. Conf. CVPR*, Jun. 2003, pp. II-264–II-271.

[16] R. Fergus, P. Perona, and A. Zisserman, "A sparse object categoty model for efficient learning and exhaustive recognition," in *Proc. IEEE Comput. Soc. Conf. CVPR*, Jun. 2005, pp. 380–387.

[17] R. Ji *et al.*, "Location discriminative vocabulary coding for mobile landmark search," *Int. J. Comput. Vis.*, vol. 96, no. 3, pp. 290–314, Feb. 2011.

[18] N. Snavely, S. M. Seitz, and R. Szeliski, "Photo tourism: Exploring photo collections in 3D," in *Proc. ACM SIGGRAPH*, 2006, pp. 835–846.

[19] R. Agrawal, T. Imielinski, and A. N. Swami, "Mining association rules between sets of items in large database," *ACM SIGMOD Rec.*, vol. 22, no. 2, pp. 207–216, Jun. 1993.

[20] D. Chen, S. Tsai, V. Chandrasekhar, G. Takacs, J. Singh, and B. Girod, "Tree histogram coding for mobile image matching," in *Proc. DCC*, 2009, pp. 143–152.

[21] V. Chandrasekhar, G. Takacs, D. Chen, S. Tsai, R. Grzeszczuk, and B. Girod, "CHoG: Compressed histogram of gradients a low bit-rate feature descriptor," in *Proc. IEEE Conf. CVPR*, Jun. 2009, pp. 2504–2511.

[22] F.-F. Li and P. Perona, "A Bayesian hierarchical model for learning natural scene categories," in *Proc. IEEE Comput. Soc. Conf. CVPR*, Jun. 2005, pp. 524–531.

[23] A. Bosch, A. Zisserman, and X. Munoz, "Scene Classification using a hybrid generative/discriminative approach," *J. IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 4, pp. 712–727, Apr. 2008.

[24] Y. Weiss, A. Torralba, and R. Fergus, "Spectral hashing," in *Proc. NIPS*, Dec. 2008.

[25] J. Yang, Y.-G. Jiang, A. G. Hauptmann, and C.-W. Ngo, "Evaluating bag-of-visual-words representations in scene classification," in *Proc. Int. Workshop MIR*, 2007, pp. 197–206.

[26] H. Jegou, M. Douze, C. Schmid, and P. Perez, "Aggregating local descriptors into a compact image representation," in *Proc. IEEE Conf. CVPR*, Jun. 2010, pp. 3304–3311.
[27] A. Gionis, P. Indyk, and R. Motwani, "Similarity search in high dimensions via hashing," in *Proc. 25th Int. Conf. VLDB*, 1999, pp. 518–529.
[28] B. Kulis and K. Grauman, "Kernelized locality-sensitive hashing for scalable image search," in *Proc. 12th ICCV*, Sep./Oct. 2009, pp. 2130–2137.
[29] H. Jegou, M. Douze, and C. Schmid, "Hamming embedding and weak geometric consistency for large scale image search," in *Proc. 10th ECCV*, Oct. 2008, pp. 304–317.
[30] J. C. van Gemert, C. J. Veenman, A. W. M. Smeulders, and J.-M. Geusebroek, "Visual word ambiguity," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 7, pp. 1271–1283, Jul. 2009.
[31] G. Csurka, C. Bray, C. Dance, and L. Fan, "Visual categorization with bags of keypoints," in *Proc. Workshop Statist. Learn. Comput. Vis., ECCV*, 2004.
[32] D. Crandall, P. Felzenszwalb, and D. Hutternlocher, "Spatial priors for part-based recognition using statistical models," in *Proc. IEEE Comput. Soc. Conf. CVPR*, Jun. 2005, pp. 10–17.
[33] F. Perronnin, C. Dance, G. Csurka, and M. Bressan, "Adapted vocabularies for generic visual categorization," in *Proc. 9th ECCV*, May 2006, pp. 464–475.
[34] J. Zhang, M. Marszalek, S. Lazebnik, and C. Schmid, "Local features and kernels for classification of texture and object categories: A comprehensive study," *Int. J. Comput. Vis.*, vol. 73, no. 2, pp. 213–238, Jun. 2007.
[35] J. Liu, Y. Yang, and M. Shah, "Learning semantic visual vocabularies using diffusion distance," in *Proc. IEEE Conf. CVPR*, Jun. 2009, pp. 461–468.
[36] J. Winn, A. Criminisi, and T. Minka, "Object categorization by learned universal visual dictionary," in *Proc. Int. IEEE ICCV*, Oct. 2005, pp. 1800–1807.
[37] D. G. Lowe, "Distinctive image features form scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, Nov. 2004.
[38] H. Bay, T. Tuytelaars, and L. Van Gool, "SURF: Speeded up robust features," in *Proc. 9th ECCV*, May 2006, pp. 404–417.
[39] Y. Ke and R. Sukthankar, "PCA-SIFT: A more distinctive representation for local image descriptors," in *Proc. IEEE Comput. Soc. Conf. CVPR*, Jun./Jul. 2004, pp. II-506–II-513.
[40] E. Eade and T. Drummond, "Unified loop closing and recovery for real time monocular SLAM," in *Proc. BMVC*, 2008.
[41] A. Irschara, C. Zach, J.-M. Frahm, and H. Bischof, "From structure-from-motion point clouds to fast location recognition," in *Proc. IEEE Conf. CVPR*, Jun. 2009, pp. 2599–2606.
[42] J. Hays and A. Efros, "IMG2GPS: Estimating geographic information from a single image," in *Proc. IEEE Conf. CVPR*, 2008, pp. 1–8.
[43] E. Kalogerakis, O. Vesselova, J. Hays, A. A. Efros, and A. Hertzmann, "Image sequence geolocation with human travel priors," in *Proc. IEEE 12th ICCV*, Sep./Oct. 2009, pp. 253–260.
[44] Y.-T. Zheng *et al.*, "Tour the world: Building a web-scale landmark recognition engine," in *Proc. IEEE Conf. CVPR*, Jun. 2009, pp. 1085–1092.
[45] R. Tibshirani, "Regression shrinkage and selection via the Lasso," *J. Roy. Statist. Soc., Ser. B*, 1997, pp. 267–288.
[46] D. Donoho, "For most large underdetermined systems of equations, the minimal l1-norm near-solution approximates the sparsest near-solution," *Commun. Pure Appl. Math.*, vol. 59, no. 7, pp. 907–934, Jul. 2006.
[47] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 2, pp. 210–227, Feb. 2009.
[48] S. A. J. Winder and M. Brown, "Learning local image descriptors," in *Proc. IEEE Comput. Soc. Conf. CVPR*, Jun. 2007, pp. 1–8.
[49] R. Ji, L.-Y. Duan, J. Chen, and W. Gao, "Towards compact topical descriptor," in *Proc. IEEE Conf. CVPR*, Jun. 2012, pp. 2925–2932.
[50] J. X. Xiao, J. N. Chen, D. Y. Yeung, and L. Quan, "Structuring visual words in 3D for arbitrary-view object localization," in *Proc. 10th ECCV*, Oct. 2008, pp. 725–737.
[51] R. Ji, X. Xie, H. Yao, and W.-Y. Ma, "Vocabulary tree incremental indexing for scalable scene recognition," in *Proc. IEEE Int. Conf. Multimedia Expo*, Jun./Apr. 2008, pp. 869–872.
[52] J. Chen, L.-Y. Duan, R. Ji, and W. Gao, "Pruning tree-structured vector quantizer towards low bit rate mobile visual search," in *Proc. IEEE ICASSP*, Mar. 2012, pp. 965–968.
[53] Y. Gao, M. Wang, Z.-J. Zha, Q. Tian, Q. Dai, and N. Zhang, "Less is more: Efficient 3-D object retrieval with query view selection," *IEEE Trans. Multimedia*, vol. 13, no. 5, pp. 1007–1018, Oct. 2011.
[54] S. Zhang, Q. Tian, G. Hua, Q. Huang, and S. Li, "Descriptive visual words and visual phrases for image applications," in *Proc. 17th ACM Int. Conf. Multimedia*, 2009, pp. 75–84.
[55] S. Zhang, Q. Huang, G. Hua, S. Jiang, W. Gao, and Q. Tian, "Building contextual visual vocabulary for large-scale image applications," in *Proc. ACM Int. Conf. Multimedia*, 2010, pp. 501–510.
[56] X. Wang, M. Yang, T. Cour, S. Zhu, K. Yu, and T. X. Han, "Contextual weighting for vocabulary tree based image retrieval," in *Proc. IEEE ICCV*, Nov. 2011, pp. 209–216.
[57] Y. Gao, M. Wang, D. Tao, R. Ji, and Q. Dai, "3-D object retrieval and recognition with hypergraph analysis," *IEEE Trans. Image Process.*, vol. 21, no. 9, pp. 4290–4303, Sep. 2012.
[58] D. M. Chen *et al.*, "City-scale landmark identification on mobile devices," in *Proc. IEEE Conf. CVPR*, Jun. 2011, pp. 737–744.
[59] Y. Gao *et al.*, "Camera constraint-free view-based 3-D object retrieval," *IEEE Trans. Image Process.*, vol. 21, no. 4, pp. 2269–2281, Apr. 2012.

Rongrong Ji (SM'14) is currently a Professor with Xiamen University, Xiamen, China, where he directs the Intelligent Multimedia Technology Laboratory and serves as a Dean Assistant with the School of Information Science and Technology. Before that, he was a Post-Doctoral Research Fellow with the Department of Electrical Engineering, Columbia University, New York, NY, USA, from 2010 to 2013. He received the Ph.D. degree in computer science from the Harbin Institute of Technology, Harbin, China. He was a Visiting Student with the University of Texas at San Antonio, San Antonio, TX, USA, a Research Assistant with Peking University, Beijing, China, and a Research Intern with Microsoft Research Asia, Beijing. He is the Guest Editor of the IEEE MULTIMEDIA MAGAZINE, *Neurocomputing*, *Signal Processing*, *ACM Multimedia Systems*, and *Multimedia Tools and Applications*, a Special Session Chair of the 2014 Indian Council of Medical Research, the 2013 IEEE Visual Communications and Image Processing Conference, the 2013 Magnetism and Magnetic Materials Conference, and the 2012 Pacific-Rim Conference on Multimedia, a Program Committee Member of more than 30 flagship international conferences, including the IEEE Conference on Computer Vision and Pattern Recognition in 2013, the International Conference on Computer Vision in 2013, and the ACM Multimedia from 2014 to 2010, and a reviewer for over 20 IEEE/ACM Transactions such as TPAMI and TKDE.

Ling-Yu Duan (M'06) received the Ph.D. degree in information technology from the University of Newcastle, Callaghan, NSW, Australia, in 2007, the M.Sc. degree in computer science from the National University of Singapore, Singapore, in 2002, and the M.Sc. degree in automation from the University of Science and Technology of China, Hefei, China, in 1999. 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. He is leading the Group of Visual Search at the Institute of Digital Media, Peking University. Since 2012, he has been the Deputy Director of the Rapid-Rich Object Search Laboratory, a joint lab between Nanyang Technological University, Singapore, and Peking University, with a vision to create the largest collection of structured domain object database in Asia and to develop rapid and rich object mobile search. He was a Research Scientist with the Institute for Infocomm Research, Singapore, from 2003 to 2008. 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. He is a member of the Association for Computing Machinery.

Jie Chen is currently pursuing the Ph.D. degree with the School of Electronics Engineering and Computer Science, Peking University, Beijing, China, where he is with 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.

Tiejun Huang (M'01) received the B.S. and M.S. degrees from the Department of Automation, Wuhan University of Technology, Wuhan, China, in 1992, and the Ph.D. degree from the School of Information Technology and Engineering, Huazhong University of Science and Technology, Wuhan, in 1999. He was a Post-Doctoral Researcher from 1999 to 2001 and a Research Faculty Member with the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China. He was also the Associated Director (from 2001 to 2003) and the Director (from 2003 to 2006) of the Research Center for Digital Media at the Graduate School, Chinese Academy of Sciences. He is currently a Professor with the National Engineering Laboratory for Video Technology, School of Electronics Engineering and Computer Science, Peking University, Beijing. His research interests include digital media technology, digital library, and digital rights management. He is a member of the Association for Computing Machinery.

Wen Gao (F'09) received the Ph.D. degree in electronics engineering from the University of Tokyo, Tokyo, Japan, in 1991. He is currently 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, from 1991 to 1995, and a Professor with the Institute of Computing Technology, Chinese Academy of Sciences, Beijing. He has served on the Editorial Boards for several journals, such as the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, the IEEE TRANSACTIONS ON MULTIMEDIA, the IEEE TRANSACTIONS ON AUTONOMOUS MENTAL DEVELOPMENT, *EURASIP Journal of Image Communications*, and *Journal of Visual Communication and Image Representation*. He has chaired a number of prestigious international conferences on multimedia and video signal processing, such as the IEEE International Conference on Multimedia and Expo and the Association for Computing Machinery's Annual Conference on Multimedia, and also served on the advisory and technical committees of numerous professional organizations.