Signal Processing: *Image Communication* 28 (2013) 368–385

Contents lists available at SciVerse ScienceDirect

Signal Processing: *Image Communication*

journal homepage: www.elsevier.com/locate/image

Learning from mobile contexts to minimize the mobile location search latency

Ling-Yu Duan [a], Rongrong Ji [a,b,*], Jie Chen [a], Hongxun Yao [b], Tiejun Huang [a], Wen Gao [a,b]

[a] *Institute of Digital Media, Peking University, Beijing 100871, China*
[b] *Visual Intelligence Laboratory, Harbin Institute of Technology, Heilongjiang 150001, China*

ARTICLE INFO

Available online 5 June 2012

Keywords:
Mobile visual search
Location recognition
Compact visual descriptor
Feature coding
Contextual learning
Benchmark dataset

ABSTRACT

We propose to learn an extremely compact visual descriptor from the mobile contexts towards low bit rate mobile location search. Our scheme combines location related side information from the mobile devices to adaptively supervise the compact visual descriptor design in a flexible manner, which is very suitable to search locations or landmarks within a bandwidth constraint wireless link. Along with the proposed compact descriptor learning, a large-scale, contextual aware mobile visual search benchmark dataset PKUBench is also introduced, which serves as the first comprehensive benchmark for the quantitative evaluation of how the cheaply available mobile contexts can help the mobile visual search systems. Our proposed contextual learning based compact descriptor has shown to outperform the existing works in terms of compression rate and retrieval effectiveness.

1. Introduction

Coming with the explosive growth of camera phones, mobile location search has received increasing focuses with a wide range of applications, for instance identifying the geographical positions of the mobile users [1–6], suggesting the touristic landmarks and photographing viewpoints [7–11], and location based advertisement [12] and augmented reality.

In a typical scenario, a mobile visual search system follows a client–server architecture. The remote server maintains a large scale database, where photos are bound with GPS or related location tags. A scalable search system with inverted indexing is offline trained for online near-duplicate visual matching, typically based on approximate nearest neighborhood search models such as hashing or bag-of-words [6,9,13–15]. In online search, the mobile user takes a query photo from his surrounding scenes, and then transmits this query to the remote server, where the corresponding location is identified based on near-duplicate matching. Subsequently, the geographical location and related information (such as augmented reality rendering results) [9,12,16] are returned to the mobile user to provide location-based services.

The query photo transmission from the mobile devices to the remote server is often over a relatively slow, bandwidth-constrained wireless link. While sending the entire query photo is often time consuming, the search experience heavily depends on how much data to be transferred. Subsequently, how to design a compact visual signature to replace the query photo delivery is one of the fundamental challenges in most existing mobile location (visual) search systems.

Related work: Coming with the ever growing computational power in the mobile devices, recent works [17–20] have proposed to directly extract compact visual

* Corresponding author at: Institute of Digital Media, Peking University, Beijing 100871, China. Tel.: +86 6462569869; fax: +86 6462569869.
E-mail addresses: lingyu@pku.edu.cn (L.-Y. Duan), jirongrong@gmail.com, rrji@pku.edu.cn, rrji@vilab.hit.edu.cn (R. Ji), cjie@pku.edu.cn (J. Chen), yhx@vilab.hit.edu.cn (H. Yao), tjhuang@pku.edu.cn (T. Huang), wgao@pku.edu.cn (W. Gao).

0923-5965/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.image.2012.05.006

descriptors on the mobile end towards a low cost wireless transmission. Comparing to the previous works on compact local descriptors *e.g.* SURF [21], GLOH [22], PCA-SIFT [23], and MSR descriptors [24], works in [17–20] focused on even lower bit rate descriptors specialized for mobile visual search.

For instance, Chandrasekhar et al. proposed a Compressed Histogram of Gradient (CHoG) [19] for compacted local feature description directly on the mobile end, which adopts both Huffman Tree and Gagie Tree to compress each local feature into approximate 60 bits. Chandrasekhar et al. also proposed to compress the SIFT descriptor with Karhunen–Loeve Transform [25], which results in approximate 2 bits per SIFT dimension. Tsai et al. [26] further proposed to transmit the spatial layouts of interest points to improve the subsequent visual matching precision. Comparing to sending the entire query photo, sending these compact local descriptors [19,25,26] is much more efficient. For instance, following the most existing local feature detector settings [22], assume there are approximate 1000 interest points detected per photo, the overall data for wireless transmission is about 8 kB, much less than sending the entire query photo (typically over 20 kB via JPEG compression).

Another feasible solution is to compress the quantized bag-of-words signatures [17,18,27]. A typical flowchart for bag-of-words based compression system is shown in Fig.1. For instance, Chen et al. proposed a Tree Histogram Coding (THC) scheme [17] to compress the sparse bag-of-words signature, which encodes the position difference of non-zero bins to achieve high compression rate. THC produced an approximate 2 kB code per image for a vocabulary with 1M words, much less than directly sending the CHoG descriptors [19] (more than 5 kB to the best of our knowledge). To maintain the visual codebook based retrieval systems in the remote server, Chen et al. [18] further compressed the inverted indices of the vocabulary tree model [13] with arithmetic coding for memory cost reduction. Recently, Ji et al. proposed to learn the compact descriptor adaptively within different geographical region [27–29,31] for mobile landmark search, which enables a region-specific, low bit rate bag-of-words landmark descriptor delivery. Successive works reported in [30] further relax the constraints of location context in compact descriptor learning, whether either unsupervised or supervised compression techniques can be deployed. Recent works in [32] also present a decent benchmark towards real-world evaluation of mobile visual search systems.

Compact visual descriptor by learning from mobile contexts: The existing works in [17–20] are based solely on the visual statistics of query photos to design compact descriptors. On the other hand, our previous work in [27] still depends on the precise segmentation of geographical regions to avoid the marginal effects in deciding the corresponding geographical region to adapt the original vocabulary maintained on the mobile end. In this paper, we further introduce a *M*ultiple-channel *C*ompact *V*isual *D*escriptor (*MCVD*) for mobile landmark (visual) search. Especially, we emphasize on a so-called adaptive channel selection mechanism, which represents the multi-tier structure of visual features towards scalable, robust, and compact query content description. Such multi-tier nature may or may not depend on the contextual (side) information. In other words, pervasive location tags, GPS coordinates, and unsupervised learning may be plugged into the compact descriptor design, allowing errors or bias in contextual tags.

The PKUBench mobile visual search benchmark: While the cheaply available mobile context such as GPS and Base Station tags are extremely beneficial to compensate visual matching, the effectiveness and robustness of such cues are left unexploited in the existing visual search benchmarks. As a result, recent works on compact visual descriptor are rarely compared among each other over a real-world, mobile visual search benchmark with rich available contextual information. Such benchmark dataset should target at real-world mobile visual search scenarios with pervasive photographing variances from phone cameras.

We believe a real-world, context rich benchmark with sufficient coverage of mobile photographing variances is important to put forward both researches and applications in mobile visual search. In this paper, we introduce the Peking University Landmarks benchmark (*PKUBench*) dataset to evaluate the context assisted mobile visual search performance. PKUBench is collected from over 198 landmarks locations in the Peking University campus, providing rich real-world photographing variances typically for mobile phone cameras. We put an extreme focus on how the cheaply available mobile context can improve the visual search performance, with a systematically designed evaluation framework to verify the robustness of mobile context in learning compact visual descriptors. We have conducted extensive experiments on PKUBench to prove that our proposed MCVD descriptor, with adaptive and error-resistant channel selection, can largely boost the mobile landmark search performance.

Paper outline: The rest of this paper is organized as follows: Section 2 briefs our Multiple-channel Descriptor Vocabulary Coding (MDVC) pipeline. Section 3 introduces our channel division/segmentation scheme based on Gaussian Mixture Model. Section 4 presents our codeword boosting scheme to learn the intra-channel compression function. Section 5 presents the details of our PKUBench dataset targeting at extensive evaluation of learning compact visual descriptor from the mobile context. Section 6 quantitatively evaluates the proposed MCVD in PKUBench, with comparisons to state-of-the-arts. We conclude our works in Section 7.

2. The multiple-channel compact visual descriptor principle

We have proposed a Multiple-channel Compact Visual Descriptor (MCVD) in this paper. MCVD is built upon our previous works of Location Discriminative Vocabulary Coding (LDVC) as presented in [27], which enables an efficient mobile landmark search even in a bandwidth-constraint wireless network. The main idea of LDVC descriptor [27] is an adaptive, asymmetric, and geographical-aware vocabulary compression, which outputs a