Noname manuscript No.
(will be inserted by the editor)

A Survey on Context-aware Mobile Visual Recognition

Weiqing Min · Shuqiang Jiang · Shuhui Wang · Ruihan Xu · Yushan Cao · Luis Herranz · Zhiqiang He

the date of receipt and acceptance should be inserted later

Abstract The phenomenal growth of the usage of mobile devices (e.g., mobile phones and tablet PCs) opens up the new service, namely mobile visual recognition, which has been widely used in many areas, such as mobile shopping and augmented reality. The rich context information (e.g., location, time and direction information) easily acquired by the mobile devices provides useful clues to facilitate mobile visual recognition, including speeding up the recognition time and improving the recognition performance. This survey focuses on recent advances in **C**ontext-**A**ware **M**obile **V**isual **R**ecognition (CAMVR) and reviews related work regarding to different context information, recog-

W. Min · S. Jiang · S. Wang · R. Xu · Luis Herranz
Key Lab of Intelligent Information Processing, Institute of Computing Technology, CAS, Beijing 100190, China
Y. Cao
Higher Education Institution Teacher Online Training Center,Beijing, China
Z. He
Lenovo Corporate Research, Beijing 100085, China

W. Min
E-mail: weiqing.min@vipl.ict.ac.cn

S. Jiang
E-mail: sqjiang@ict.ac.cn

S. Wang
E-mail: wangshuhui@ict.ac.cn

R. Xu
E-mail: rhxu@ict.ac.cn

Yushan Cao
E-mail: caoyushan@enet.edu.cn

Luis Herranz
E-mail: luis.herranz@vipl.ict.ac.cn

Zhiqiang He
E-mail: lirong2@lenovo.com

This is a pre-print of an article published in Multimedia Systems. The final authenticated version is available online at: https://doi.org/10.1007/s00530-016-0523-8

nition method, recognition type and various application scenarios. Finally, we discuss the future research directions in this field.

Keywords mobile visual recognition · context · survey

1 Introduction

Recent years have witnessed an explosive growth in the use of mobile devices. Built-in cameras and network connectivity make it increasingly appealing for users to snap pictures of objects and then obtain relevant information about the captured objects, which is referred as the mobile visual recognition. For example, a user takes a photo of a landmark and obtains the returned textual information (e.g., the landmark tags and relevant descriptions), related images (e.g., different views of the same landmark) or 3D model [62] about the landmark. Mobile visual recognition is particularly useful in applications such as mobile shopping [31, 58], mobile landmark recognition for tourists [8], and mobile location recognition for augmented reality [83]. Furthermore, such mobile visual recognition functionalities have been shown in many commercial systems, such as Google "Goggles"[1], Amazon "Snaptell"[2], and "Kooaba"[3].

A lot of work [96, 26, 32, 10] on mobile visual recognition is directly to extract the visual features for image representation. Recently, the deep learning technique [42] has achieved great progress for the feature representation. Because of the low storage limitation of mobile devices, the most popular way is to compress the visual features on the mobile side by some coding method, such as SURF [6], CHoG [7] and BoHB [31]. However, one shortcoming of these work is that these methods are mainly based on the content analysis alone, and do not consider the rich context information (e.g., the GPS and time information) easily acquired by the mobile device, which can be useful for mobile visual recognition.

In fact, mobile equipments brings a lot of context information, which can be categorized into two levels: One is the internal context information which is intrinsically contained in the mobile devices such as stored textual/visual content, camera and other sensor's parameters. The other is the external context information which could be easily acquired by the mobile device such as time and geolocation. Researchers have exploited many of them to improve the recognition performance. More common used contexts include location, direction, time, text, gravity, acceleration, and other camera parameters. For example, in [84], the content analysis is essentially filtered by a predefined area centered at the GPS location of the query image. Chen *et al.* [8] utilized the GPS information to narrow the search space for landmark recognition. Ji *et al.* [41] designed a GPS based location discriminative vocabulary coding scheme, which achieves extremely low bit rate query transmission for mobile landmark search. Chen *et al.* [19, 16] combined the visual information with the context information including the location and the direction information for mobile landmark recognition. Runge *et al.* [74] suggested the tags of images using the location name and time period. Gui *et al.* [29] addressed mobile scene recognition by fusing outputs of inertial sensors and computer vision techniques. In such cases, utilizing the context information in mo-

[1] www.google.com/mobile/goggles/
[2] http://www.snaptell.com/.
[3] www.kooaba.com

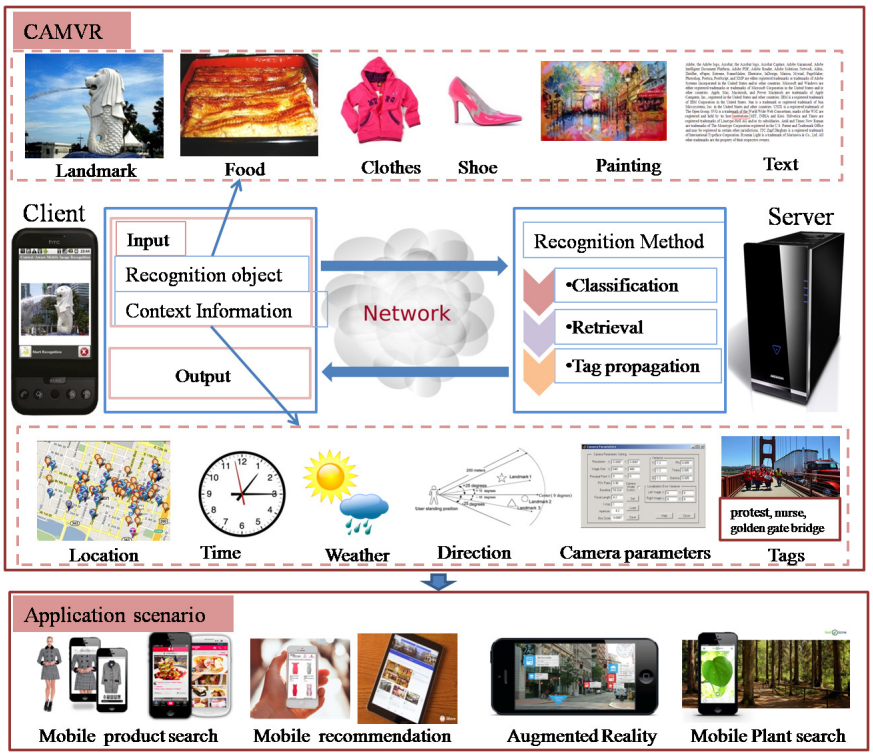

Fig. 1 An overview of CAMVR. Top: the overflow of CAMVR; Bottom: application scenarios of CAMVR

bile visual recognition can speed up the recognition time and improve the recognition performance.

In this survey, we give a comprehensive overview of **C**ontext-**A**ware **M**obile **V**isual **R**ecognition(CAMVR). A typical pipeline for CAMVR is shown in the top of Fig.1. For the client side, the input is the captured object (e.g., one landmark, food, clothes and painting) and the context information acquired by the mobile phone (e.g., location, time and weather). After the input information is sent to the server, one recognition method (e.g., classification and retrieval) from the server side is selected to recognize the object and the relevant information is returned to the user as the output. From the overall system, we can review CAMVR from three different aspects, namely context information, recognition method, recognition types. Based on the CAMVR system, there are great potential applications (in the bottom of Fig.1), such as mobile product search, mobile recommendation, and augmented reality.

The rest of the survey is organized as follows: In Sections 2 through 4, we survey the state-of-the-art approaches of CAMVR according to different context information, different recognition methods and different recognition types, respectively. In Section 5, we introduce various application scenarios based on CAMVR. Finally, we conclude the paper with a discussion of future research directions in Section 6. $i_l^{'} = \sigma(W_{ix}x_l + W_{im}m_{l-1} + W_{iq}g)$ $f_l^{'} = \sigma(W_{fx}x_l + W_{fm}m_{l-1} + W_{fq}g)$ $o_l^{'} = \sigma(W_{ox}x_l + W_{om}m_{l-1} + W_{oq}g)$ $c_l^{'} = f_l^{'} \odot c_{l-1}^{'} + i_l^{'} \odot h(W_{cx}x_l + W_{cm}m_{l-1} + W_{cq}g)$ $m_l = o_l^{'}c_l^{'}$

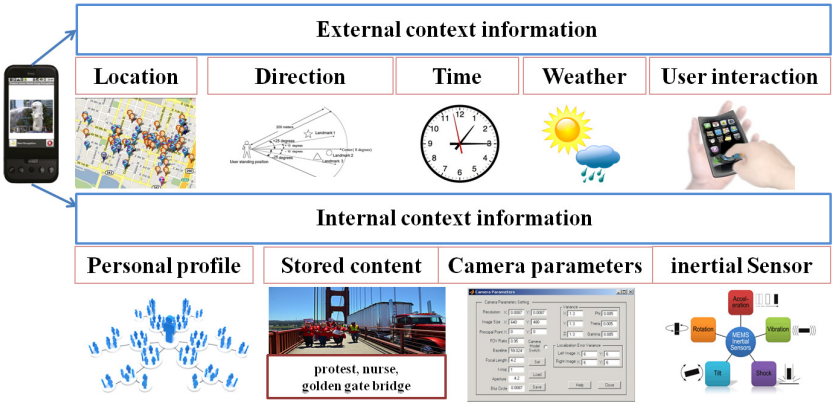

Fig. 2 Different kinds of context information

2 Context information

In this section, we review related work on CAMVR based on different types of context information. As shown in Fig. 2, the context information can be divided into two levels: One is the external contextual information which could be easily acquired by the mobile devices such as location and time. The other is the internal contextual information which is intrinsically contained in the mobile devices such as personal profile (e.g., user's relationship), stored textual/visual content, and camera's parameters. CAMVR can exploit various forms of context information to facilitate recognition. For example, if the location information is available, the system can significantly reduce the search scope for the captured object, which in turn will greatly improve recognition accuracy and speed [8]. Direction refers to the shooting direction, which is an necessary compliment for location information, especially for recognizing faraway target or scene [19]. As lots of images, especially those uploaded to social networks, are with text descriptions or tags input by users, text contexts play an important role in recognizing uploaded images [1].

Among all context information, location is the most common context information for visual recognition. Therefore, we divided the use of context into the following three groups, as shown in Table 1:

- Location. This group of work [47, 97] only use the location information to improve the recognition recognition.
- Context with location. Besides the location information, this group of work [15, 16] also employ other context information, such as time and direction, which are probably complimentary for location information
- Context without location. This group of work [29] does not use the location information, but other context information (e.g., the inertial sensors' parameters) for CAMVR.

Table 1 Summarization of CAMVR based on different types of context information

Context information	**Location**	**Context with location**	**Context without location**
Representative work	Tsai *et al.* [84] Fritz *et al.* [25] Quack *et al.* [72] Takacs *et al.* [83] Zhu *et al.* [100] Yap *et al.* [93] Chen *et al.* [8] Ji *et al.* [39] Ji *et al.* [41] Duan *et al.* [23]	Benjamin *et al.*[70] Naaman *et al.* [64] Sinha *et al.* [81] Chen *et al.* [18] Li *et al.* [54] Chen *et al.*[15] Chen *et al.* [19] Guan *et al.* [28] Guan *et al.* [27]	Li *et al.* [49] [Xia *et al.* [88] Zhang *et al.* [99] Gui *et al.* [29] Hao *et al.* [30] Qin *et al.* [71] You*et al.* [94]

2.1 Location

Nowadays, mobile devices are widely equipped with embedded GPS chips. As a result, visual data associated with geographical or location tags can be easily produced in our daily lives. With the help of available location information, the mobile visual recognition system can significantly reduce the search scope for the captured object, which in turn will speed up the recognition time and improve the recognition accuracy [93]. For example, Takacs *et al.* [83] used the GPS signal to retrieve only images falling in nearby location cells. Amlacher *et al.* [2] exploited the GPS information to narrow the search space for mobile object recognition. Similar to [83,2], Kuo *et al.* [47] also introduced the GPS constraints in the retrieval process on inverted indexing so that they can achieve a real-time image retrieval system. In [25,84,66], the GPS location information is also utilized to assist in content-based mobile image recognition. With the aid of the location information, the challenge in differentiating similar images that are captured in different area can be reduced substantially. Xie *et al.* [90] used a multimodal search scheme which uses the image content and user location to increase the search accuracy while Zhu *et al.* [100] used multi-modality clustering of both content and GPS information for efficient image management and search. Compared with the work based on the combination between visual information and GPS information, Zamir *et al.* [97] proposed a multimodal approach which incorporates the location information, business directories, textual information, and web images in a unified framework to identify businesses in an image. In addition, Mai *et al.* [60] combined the GPS information and 3-D model to match the query image.

In addition to using the GPS information in general visual recognition tasks, a lot of work [72,93,46,8,39,38,41,23,98,82,33,91] focuses on utilizing the GPS information for specific tasks, such as roadside sign recognition [79], mobile landmark recognition [72, 93,46,8,39,38,41,23,51] and mobile food recognition [82,33,91]. For example, Seifert *et al.* [79] proposed a mobile system based on a GPS sensor for roadside sign localization and classification. Chen *et al.* [8] utilized the GPS coordinates to narrow the search space for landmark recognition. Ji *et al.* [39,38,41] designed a GPS based location discriminative vocabulary coding scheme, which achieves extremely low bit rate query transmission for mobile landmark search. Song *et al.* [82] introduces the geo-constraints for food image recognition.

However, GPS-based mobile visual recognition has some drawbacks [8,57,78] that make it non-ideal in real applications: Firstly, the embedded GPS modules rely on a satellite navigation system and need at least four satellites to provide sufficient posi-

tioning accuracy. As a result, the estimated GPS location in a crowded urban scene or on a cloudy day is error prone, usually leading to an error of 50∼100 meters. The large GPS error of the captured image will result in wrong recognition. Secondly, besides the GPS information, there are other context information available from the mobile devices. The effective integration of different context information will further improve the recognition performance. Therefore, some work [8, 57] has resorted to combining the GPS information with other context information (e.g., direction information) for enhance the recognition performance.

2.2 Context with location

Besides the GPS information, other context information, such as direction and time information is also easily acquired by mobile devices equipped with digital compass and other sensors. Combining the content information with richer context information will facilitate the recognition performance. For example, Benjamin *et al.* [70] present a system iPiccer to infer photo tags from its location and orientation. Chen *et al.* [17, 18, 15, 12, 19, 16, 14] incorporated the location and direction information to perform mobile landmark recognition. The direction information is obtained through the built-in digital compass of mobile devices and is complementary to the location information. Similarly, Li *et al.* [50] proposed a boosting algorithm to integrate the visual content and two types of context information, including the location and direction for mobile landmark recognition. Tao *et al.* [28, 27] implemented a GPS-based and heading-aware RankBoost algorithm to reduce the dimensionality of the bag-of-features(BOF) descriptors for mobile location recognition. In addition, the location and time are also often combined in mobile visual recognition. For example, Yang *et al.* [48, 92] proposed to utilize the geographic location and date/time where the photo was taken to create automatic spatial and temporal indexes for image retrieval. Lin *et al.* [54] generated tags for content from metadata, which is pre-filtered based on the location and time information. Runge *et al.* [74] suggested the tags of images using the location name and time period.

Furthermore, the integration of more than two kinds of context information with location information is also utilized for mobile visual recognition. Ahern *et al.* [1] proposed a system for the media annotation via various context information, including restaurants, events, venues near the user's location, past tags from the user and the user's social network. Naaman *et al.* [64] balanced all the tag sources to generate a prioritized suggested tag list by using several context information, including the location, the tags' social and temporal context. Li *et al.* [52] utilized three types of context information, namely location, user interaction and Web for mobile image annotation. Huang *et al.* [35] utilized the clustering and similarity-based approaches for photo tagging using various context information, such as date, time, location, environment noise, and human faces. Wang *et al.* [81] conducted photo annotation by exploiting the following four kinds of meta information: a. optical meta layer, which contains the metadata related to the optics of the camera, e.g., the focal length and exposure time; b. temporal meta layer, which contains the time stamp of the instant where the photo was taken; c. spatial meta layer, which contains the spatial coordinates of the places where pictures were shot; and d. human induced meta layer, which contains the tags and comments posted by people.

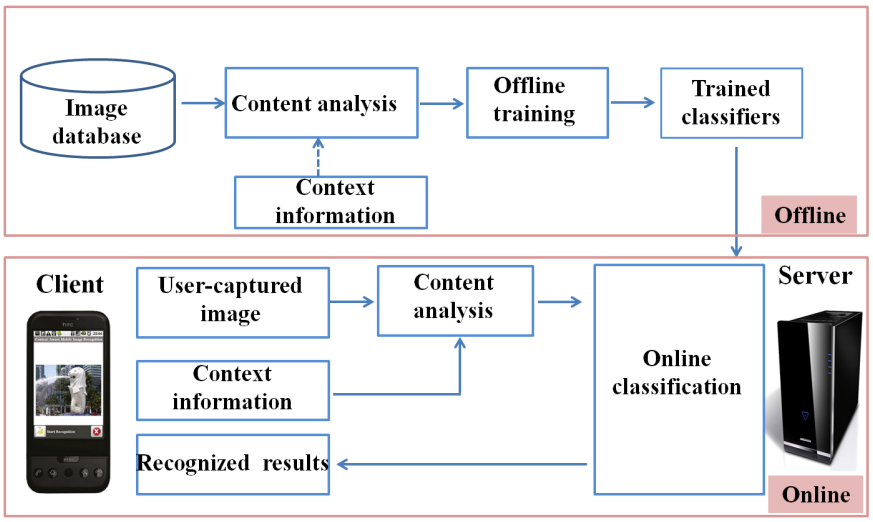

Fig. 3 Mobile visual classification system overview

2.3 Context without location

Some specific mobile applications do not need the location information but other context information, such as camera and other sensors' parameters. For example, Gui *et al.* [29] fused outputs of inertial sensors and computer vision techniques to address the mobile scene recognition problem. Hao *et al.* [30] proposed a novel technique for point of interest detection from sensor-rich videos by leveraging sensor-generated meta-data (camera locations and viewing directions). Pei *et al.* [69] studied the viewing angle estimation by exploiting the visual appearance of the query, which can be further improved by incorporating the coarse mobile context including gyro or compass information. Xia *et al.* [88] proposed an effective and efficient geometric context-preserving progressive transmission method for mobile visual search. Qin *et al.* [71] proposed a mobile phone-based collaborative system TagSense that senses the people, activity, and context in a picture, and merges them carefully to create tags on-the-fly. In addition, some work such as [49, 94, 99, 76] considered the user interaction as the context information for mobile image retrieval.

3 Recognition methods

Existing recognition methods of CAMVR can be summarized into the following three categories:1) Classification based method; 2) Retrieval based method and 3) Tag propagation based method. Table 2 summarizes the representative work for each kind of method.

3.1 Classification based method

Classification based method firstly trains a recognizer for each object (e.g., landmark and food) by integrating the content and context analysis, and then recognizes the

query image using the trained classifer and the context information associated with the query image. Figure 3 gives an overview of a classification based CAMVR system, consisting of content analysis (extracting features from the image), context information extraction (for example, determining the location through GPS), and classification (identifying which category the captured object belongs to). In the following sections, we'll present the state-of-art content analysis with context information and classification algorithms, respectively.

Table 2 Summarization of CAMVR based on different recognition methods

Recognition method	**Classification**	**Retrieval**	**Tag propagation**
Representative work	Fritz *et al.* [25] Lim *et al.* [53] Chen *et al.* [13] Chen *et al.* [17] Li *et al.* [50] Chen*et al.* [19]	Girod *et al.* [26] Yu *et al.* [96] He *et al.* [31]	Naaman *et al.* [65] Ahern *et al.* [1] Arandjelović *et al.* [4] Li *et al.* [52]

3.1.1 Content analysis with context information

Content analysis is mainly to extract features from the image. We can broadly categorize the visual features into global and local features [93]. Global features characterize an image's overall properties and only describe the image's global statistical properties, ignoring regions of interest. Therefore, most mobile visual recognition systems use local features, which aim to represent the image content using local features extracted from salient regions or patches within the image. The local features [16] can be divided into two classes:1) local patch image representation [53, 13] that uses visual features extracted from the local patches in the image for recognition and 2) bag-of-words (BoW) histogram representation [25, 86, 18, 96, 31] that generates a BoW histogram for each image through vector quantization.

For local patch image representation, Lim *et al.* [53] employed a discriminative patch selection algorithm that extracts the most discriminative patches from an image. It uses the patch density likelihood ratio to find discriminative patches. However, this method often leads to a high false-positive rate. In order to solve this problem, Chen *et al.* [13] firstly extracted a set of multi-scale patches of images and then selected discriminative pathes based on a Gaussian mixture model. The dense multi-scale patch representation is used to ensure that the extracted features are more robust towards changes in scale of the landmarks. However, compared with the local patch image representation, BoW generally requires less computational time because the image descriptor is in the form of a codeword histogram, which usually has 200 to 600 dimensions. In view of mobile devices' real-time requirements, BoW is suitable for mobile landmark recognition.

The SIFT descriptor [59] is one of the most widely used BoWs in state-of-art mobile visual recognition [25, 49, 8, 96]. However, conventional SIFT involves detection of a large number of salient keypoints and extraction of a 128-dimensional feature vector centered on each of these keypoints. To reduce standard SIFT's computational cost, researchers have proposed several SIFT-based variants that reduce the number of keypoints and the feature vector's dimensions, such as clustering to group similar

keypoints [49] and Informative-SIFT [25], SURF [20,83] and CHoG [7]. Recently, He *et al.* [31] proposed "Bag of Hash Bits" (BoHB), in which each local feature is encoded to a very small number of hash bits and it significantly outperforms CHoG in mobile product search.

Integrating the context information into the content analysis can make the features more discriminative. ***Integration between content and context analysis*** is that the context information is not only used in the online phase, but also offline feature learning. For CAMVR, most work [8,18,17,19,16] mainly use the context information to reduce the search space for the query image in the online phase. During the offline learning phase, the context information is not incorporated. However, some work [41, 50,15,14] also fully utilized the context information in the offline discriminative feature learning. For example, Ji *et al.* [41] not only used the GPS data for image filtering, but also incorporated the GPS information into the TF-IDF scheme to weight various visual words to build a location discriminative vocabulary and further improved the landmark search performance. Li *et al.* [50] improved the content-based recognition performance by incorporating recognition results from various context-based vocabulary trees (VTs) built upon location and direction context information. Similar to [50], Chen *et al.* [15, 14] exploit both location and direction information to learn a discriminative compact vocabulary (DCV). Xu *et al.* [33,91,82] combined the GPS information and visual information for discriminative training and food recognition.

3.1.2 Classification algorithms

Most work [53,13,17,19,18] employs Support Vector Machines (SVMs), which is a state-of-the-art algorithm that can be very fast at the testing step, while demonstrating exceptional generalization ability. For example, Chen *et al.* [18] employed multi-class support vector machine (SVM) with the new spatial pyramid kernel (SPK) to train the landmark classifiers. In [16], they proposed to employ ensemble of classifiers using fuzzy support vector for training. Xu *et al.* [91] adopted the location-adaptive SVM classification for training. Li *et al.* [50] used a multi-class Adaboost classifier, which constructs a strong classifier by combining weak classifiers. Combined with properly extracted image features, these discriminative classification methods can perform well in the presence of background clutter, viewpoint changes, and partial occlusions. In addition, Fritz *et al.* [25] applied the MAP classification to mobile visual applications.

3.2 Retrieval based method

Compared with classification based method, feature representation from retrieval based method is similar to classification based method. The difference is that classification based method is to train the model on a training set and use the trained model to conduct recognition, while retrieval based method is that to get the "closest" image to the input image from database by feature matching algorithm and then assign the closest label to the input image. Fig. 4 shows an overview of a retrieval based CAMVR system, consisting of content analysis , context information extraction, image matching, and geometrical verification. Since content analysis and context information extraction is similar to the classification based method, we mainly review the approaches of the feature matching and geometric verification (GV) algorithm [26,83]. For general methods, feature matching finds a small set of images in the database that have many

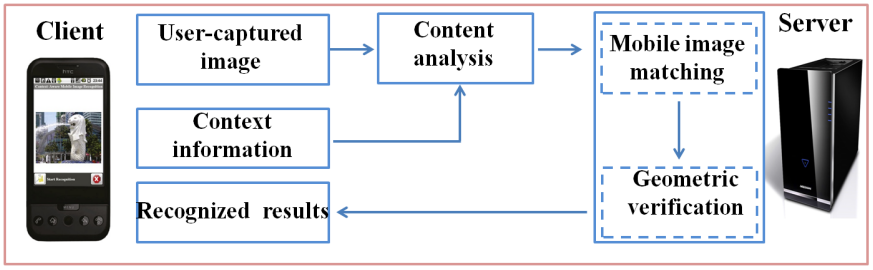

Fig. 4 Mobile visual retrieval system overview

features in common with the query image, and the GV step rejects all matches with feature locations that cannot be plausibly explained by a change in viewing position [26]. There are also some different kinds of image matching strategies. For example, Yu *et al.* [96] present a mobile application that can teach mobile users to capture pictures that can distinctively represent the surrounding scenes. Besides feature matching with hash bits and geometry verification, He *et al.* [31] further introduced the boundary reranking algorithm to improve the retrieval performance.

3.3 Tag propagation based method

Tag propagation based method [36,34,4,5] is to annotate the query image by propagating the tags of other images to this image, where images with tags are similar to this image based on the content or context information. Arandjelović *et al.* [4] (Fig.5) is firstly to visually match the sculpture image to get visually similar images, and then propagate the tags of these images to the sculpture to name it. In contrast, some work resort to context information to restrict the tag propagation process for visual annotation. For example, Naaman *et al.* [65] assigned a label to a new photo by propagating the labels of the photos taken within the same location. Ahern *et al.* [1] proposed a mobile system ZoneTag to support media annotation via context-based tag suggestions. Sources for tag suggestions include past tags from the user and other context information. Li *et al.* [52] utilized the context information, such as the location information, direction information, time information, domain information (e.g.interaction between user and information server) and web information to restrict the tag propagation process for image annotation. In addition, they also considered different tag distributions at different places in propagating tags to the query images.

4 Recognition types

The mobile recognition types can be generally categorized into the following three groups: 1) mobile location recognition (e.g., mobile landmark recognition), 2) mobile product recognition (e.g., mobile food recognition and mobile clothes retrieval), and 3) other mobile object recognition (e.g., mobile painting recognition and mobile document recognition). Table 3 summaries representative work for each kind of recognition type.

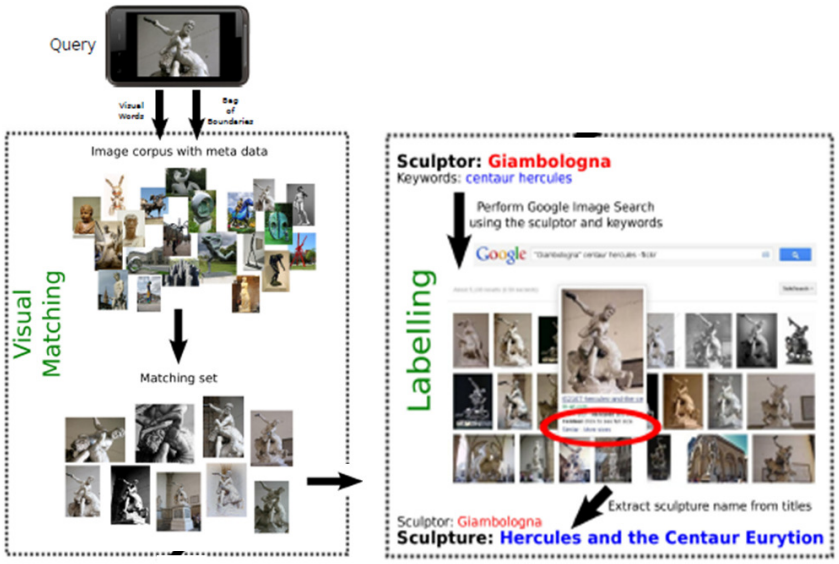

Fig. 5 overview of mobile sculpture annotation [4]

Table 3 Summarization of CAMVR based on different recognition types

Recognition type	Location	Product	Other objects
Representative work	Schroth *et al.* [78] Girod *et al.* [26] Yu *et al.* [96,95] Liu *et al.* [57,56] Duan *et al.* [23] Guan *et al.* [27] Liu *et al.* [55] Chen *et al.* [8] Ji *et al.* [41] Chen *et al.* [17,18] Chen *et al.* [15] Chen *et al.* [19,16] Min *et al.* [62] Zhang *et al.* [98]	Tsai *et al.* [85] He *et al.* [32,31] Shen *et al.* [80] Maruyama *et al.* [61] Kawano *et al.* [43] Liu *et al.* [58] Di *et al.* [22]	Auack *et al.* [72] Gui *et al.* [29] Mouine *et al.* [63] Duan *et al.* 2014[24]

4.1 Mobile location recognition

Most interesting Location Based Services (LBSs) could be provided in densely populated environments, which include urban canyons and indoor scenarios [78]. However, GPS is hardly available in these urban canyons and indoors. Visual location recognition enables LBSs in these densely populated areas without the need for complex infrastructure. Therefore, mobile visual location recognition offers a service complementary to GPS or network based localization.

Schroth *et al.* [78] focused on solving the latency and limited storage capacity of mobile devices in mobile location recognition. Girod *et al.* [26] proposed to use compact feature descriptors and spatial coding schemes for mobile visual searching, which

also proves very useful for vision-based mobile localization. Yu *et al.* [96,95] present a mobile location search application that can teach mobile users to capture pictures that can distinctively represent surrounding scenes. Duan *et al.* [23] proposed to learn an extremely compact visual descriptor from the mobile contexts towards low bit rate mobile location search. Tao *et al.* [28,27] proposed a memory-and computation-efficient encoding algorithm to enable efficient on-device mobile visual location recognition. Schroth *et al.* [77,68] partitioned the whole work space into overlapped sub regions and design a prior knowledge (such as Cell-ID) based strategy to download the visual words and associated inverted file entries in an incremental way to perform location recognition directly on mobile devices. Unlike these systems, Liu *et al.* [57,56,55] proposed framework aims to provide complete geo-context scene information: location, viewing direction, and distance to the captured scene with a higher accuracy than using only the GPS function. Such accurate geo-context can lead to a better experience of LBSs for mobile users.

In mobile visual location recognition, mobile landmark recognition which uses the camera phone to capture a landmark and find out its related information, is receiving more and more users' attention for its great potentials in travel recommendation. Chen *et al.* [8] built up a million-scale street view image dataset available to the public and conduct concrete experiments to evaluate their landmark retrieval scheme. Ji *et al.* [40,41] proposed a discriminative vocabulary coding scheme for mobile landmark search. similar to [40,41], Zhang *et al.* [98] also proposed to learn a geo-discriminative codebook for mobile landmark recognition. Besides the location information, Chen *et al.* [17,18,15,12,19,16,14] also incorporated the direction information to perform mobile landmark recognition. Similarly, Li *et al.* [50] also used the two types of mobile context: location and direction information for mobile landmark recogntion. Different from these work, Min *et al.* [62] proposed a robust 3D model based method to recognize query images with corresponding landmarks. The proposed search approach starts from a 2D compressed image query input and ends with a 3D model search result.

4.2 Mobile Product Recognition

Mobile product search is one of the most popular mobile search applications, because of the commercial importance and wide user demands. Tsai *et al.* [85,26] present a fast and scalable mobile product recognition system for the camera-phone, where the database primarily comprises products packaged in rigid boxes with printed labels, such as CDs, DVDs, and books. He *et al.* [32,31] encoded each local feature into a very small number of hash bits for efficient mobile product search on different product dataset, which are crawled from online shopping companies, such as Ebay.com, Zappos.com, and Amazon.com. Shen *et al.* [80] proposed to simultaneously retrieve visually similar product images, and localize/identify the product instance in the query image for mobile product images retrieval.

Among all products, mobile food recognition and mobile clothes recognition are particularly useful for great business potentials. For ***mobile food recognition***, Maruyama *et al.* [61] proposed a system which extracts the color features and recognizes 30 kinds of food ingredients on a mobile device. Kawano *et al.* [43] proposed a real-time food recognition system, where a user firstly draws bounding boxes by touching the screen, and then the system starts food item recognition within the indicated bounding boxes. Kawano *et al.* [45,44] computed Fisher vectors over HOG patches to develop a real-time

mobile food recognition system on a larger food dataset. Oliveira *et al.* [67] presented a semi-automatic system to recognize prepared meals which is light weight and can be easily embedded on a camera-equipped mobile device. Different from these work, Xu *et al.* [91] proposed a framework incorporating discriminative classification in geolocalized settings and introduce the concept of geolocalized models for food recognition. As for ***mobile clothes recognition***, Liu *et al.* [58] proposed a "street-to-shop" clothing retrieval model, where a user takes a photo of any person, then similar clothing from online shops are retrieved using the proposed cross-scenario image retrieval solution to facilitate online clothing shopping. However, this system focuses on recognition or retrieval at the category level (e.g. suit, dress, sweater). Di *et al.* [22] proposed a fine-grained learning model and multimedia retrieval framework to extract and match different attributes for clothing style recognition and retrieval. Cushen *et al.* [21] presented a mobile visual clothing search system whereby a user can either choose a social networking photo or take a new photo of a person wearing clothing of interest and search for similar clothing in a retail database. The GPS information is used to re-rank results by retail store location.

4.3 Other mobile object Recognition

In order to integrate mobile visual search techniques into digital library, Duan *et al.* [24] proposed a novel mobile document image retrieval framework. Ruf *et al.* [73] recognized paintings in art galleries for mobile museum guide. Gui *et al.* [29] addressed the recognition of large-scale outdoor scenes on smart-phones by fusing outputs of inertial sensors and computer vision techniques. Mouine *et al.* [63] designed a mobile plant recognition system for plant identification. Auack *et al.* [72] identified an object from a query image through multiple recognition stages, including local visual features, global geometry, and optionally also metadata such as GPS location.

5 Application scenarios

There are increasing amount of applications related to CAMVR, which can be briefly categorized into the following six groups: 1) mobile search, 2) mobile recommendation, 3) mobile shopping, 4) mobile navigation, 5) mobile augmented reality and 6) other potential mobile applications.

5.1 Mobile search

Mobile search has made a great contribution to the market. According to a leading market research firm eMarketer[4], by 2011, mobile search will account for around $715 million, or almost 15% of a total mobile advertising market worth nearly $4.7 billion. As one important branch of mobile search, mobile visual search is particularly useful. There have been many commercial systems on mobile product search such as Google "Goggles" , Amazon "Flow"[5], "Kooaba", and Nokia "Point and Find"[6]. Google Goggles is

[4] https:/en.wikipedia.orgwikiMobile_search
[5] http://flow.a9.com
[6] pointandfind.nokia.com

a mobile application that lets users search the web using pictures (e.g., books, artworks and wine) taken from their mobile phones. Amazon Flow let users snap a photo of the cover of any CD, DVD, book, or video game, and the application will automatically identify the product and find ratings and pricing information online. Kooaba receives a snapped image as the query and displays related information, further links and available files, applied to wine lists, printed catalogues, etc. Point and Find is a service offered by Nokia that uses visual search technology to let users find more information about the surrounding objects, places, etc. In addition, there is a lot of work on mobile product search [85, 26, 31] in academy, such as mobile food recognition [43, 45] and mobile clothes retrieval [58, 22, 21]. In addition, some work [17, 18, 15, 12, 19, 16, 14] incorporated the location and direction information to perform mobile landmark search.

5.2 Mobile recommendation

The recognized results can be used for mobile recommendation. For example, Mobile landmark recognition [8, 16, 51] can be further used for travel recommendation. Maruyama *et al.* [61] proposed a mobile cooking recipe recommendation system by recognizing food ingredients such as vegetables and meats. Zhang *et al.* [99] allowed a mobile user to take a photo and naturally indicate an object-of-interest within the photo via circle based gesture called gesture. Both selected object-of-interest region as well as surrounding visual context in photo are used in achieving a search-based recognition by retrieving similar images. Consequently, social activities such as visiting contextually relevant entities (i.e., local businesses) are recommended to the users based on their visual queries and GPS location. Viana *et al.* [87] (Fig.6) proposed a mobile photo and video recommendation system, which including acquiring the user's context, enriching and annotating the context data, performing a similarity analysis, and providing photo recommendations.

5.3 Mobile shopping

The wide use of mobile devices leads to the fast development of mobile shopping. There are some commercial systems for mobile shopping based on CAMVR. For example, oMoby[7] offers a shopping service that helps users find information about products by snapping a photo, such as links to retailers offering product information, reviews, prices, and more. Visual Fashion Finder provided by Cortexica Vision Systems allows consumers to take a picture of an item of clothing or fashion accessory with a mobile device, and automatically finds similar items from a database of inventory. In academy, some work, such as [58] proposed a mobile clothes retrieval model: a user takes a photo of any person, then similar clothing from online shops are retrieved using the proposed cross-scenario image retrieval solution to facilitate online clothing shopping. Di *et al.* [22] proposed an attribute-based search and retrieval schema for mobile clothing shopping, which has multiple potential mobile applications including style-based retrieval and navigation, as well as automatic style tagging for query images. Cushen *et al.* [21] presented a mobile visual clothing search system, which allows that a user can either choose a social networking photo or take a new photo of a person wearing clothing of

[7] www.omoby.com

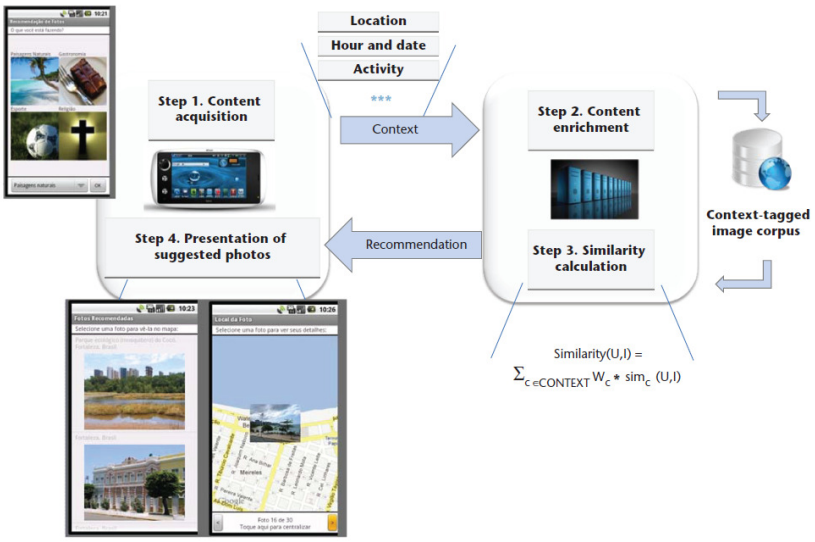

Fig. 6 overview of mobile photo and video recommendation system [87]

interest and search for similar clothing in a retail database. The GPS information is introduced to re-rank results by retail store location. You *et al.* [94] focused on improving visual search based mobile shopping experience by using machine and crowd intelligence, where the user interaction can be considered as the context information.

5.4 Mobile navigation

Mobile visual location search [78] and mobile landmark recognition [17,18,15,12,19, 16,14] can be used for mobile navigation. Je *et al.* [37] introduced the street searching service for mobile navigation. As shown in Fig.7, the buildings around crossroads are appropriate for the image based localization. Therefore, as the first step, a user takes a photo of one landmark around the crossroads. The query photo is then transmitted to the searching server. In the second step, the user receives the location and he or she is asked to determine which direction is to be navigated. In the third step, the user looks around the selected direction with traditional map and multi-perspective panoramic street views. It can help us search and find out somewhere more intuitively. In addition, Liu *et al.* [57,56] present a novel approach to mobile visual localization that accurately senses geographic scene context according to the current image (typically associated with a rough GPS position) and applied it to mobile navigation.

5.5 Mobile augmented reality

Mobile augmented reality (MAR)[9] is a wide class of applications where the mobile devices augment users' perception of the world. MAR processes a stream of viewfinder frames captured by a mobile device's camera to recognize [26], track, and augment objects that appear in these frames. Chen *et al.* [11] streamed live videos on the mobile

Fig. 7 overview of mobile visual location search for navigation [37]

phone to the remote server, on which a SURF-based recognition engine was used to obtain features. In their latest work [9], they developed a new methods for interframe coding of a continuous stream of global signatures that can reduce the bitrate by nearly two orders of magnitude compared to independent coding of these global signatures, while achieving the same or better image retrieval accuracy.

5.6 Other potential mobile applications

Mobile visual recognition can also be used for product placement[8]. For example, users can snap a picture of a poster of a popular Bollywood movie and instantly be connected with more content such as movie trailers, and Tweets from the film's actors. The technology offers opportunities for new partnerships involving product placement, in which users can see a product, snap a picture and purchase it online via the mobile device at the moment of intent. In addition, the mobile visual recognition can also be used in online communication and intelligent interaction.

6 Conclusions and future research directions

In this survey, we have reviewed the recent work on context-aware mobile visual recognition (CAMVR). We firstly introduced the available mobile contexts which are common used, and showed that the location context is popular for various recognition tasks,

[8] http://marketingland.com/mobile-visual-search-begins-bridge-gap-real-digital-world-101673

and other types of contexts are often used as complementary. Then we described different recognition methods, and showed that most work are based on classification or retrieval. Next, we listed different recognition types. Finally, we categorized the application scenarios, which showed a promising prospect for CAMVR.

Although tremendous progress has been made, there are still several open issues that need to be addressed in future work, including: 1) how to combine more context information; 2) how to design compact and discriminative descriptors; 3) how to effectively integrate content and context information; and 4) how to consider user's intention.

Firstly, compared with general mobile visual recognition, one goal of CAMVR is to utilize rich context information to speed up the recognition time and improve the recognition performance. However, the context information of most existing work [100] [93] [8][39] [41] [23][18,15,19,14] is limited to GPS information or two kinds of context information. The constraint of more context information can further speeded up the recognition time and thus the real-time requirement of mobile visual recognition is more easily satisfied. Therefore, the method of effectively combining more context information [64,35] is desirable.

Secondly, limited storage capacity and real-time requirement are two limitations of mobile visual recognition. This limited the use of very high-dimensional feature descriptors. Therefore, the smaller descriptors with comparable discriminative performance are needed. Although some work [7,31] have designed compressed descriptors, which achieves almost identical performance as common SIFT descriptors. However, they still do not satisfy the requirement of some applications, such as mobile augmented reality. Therefore, designing compact and discriminative feature descriptors ought to be studied.

Thirdly, most work [8,18,17,19,16] mainly use the context information to reduce the search space for the query image in the online phase. However, during the offline learning phase, the effective combination between the content and context information probably improve the recognition performance. Some work [41,15] utilized the context information (e.g., GPS and direction information) in the offline discriminative feature learning and improved the recognition performance. Therefore, it is interesting to integrate more context information to the content information for more effective feature learning in the offline phase.

Finally, since mobile visual recognition is to satisfy user's needs, the ideal mobile visual recognition should take the user intent into account. Few work [100] consider the user intent in mobile visual recognition. Therefore, how to incorporate the user intent into mobile visual recognition is probably an interesting research topic. [75,89,3]

References

1. Ahern, S., Davis, M., Eckles, D., King, S., Naaman, M., Nair, R., Spasojevic, M., Yang, J.: Zonetag: Designing context-aware mobile media capture to increase participation. In: Proceedings of the Pervasive Image Capture and Sharing, 8th Int. Conf. on Ubiquitous Computing, California (2006)
2. Amlacher, K., Paletta, L.: Geo-indexed object recognition for mobile vision tasks. In: Proceedings of the 10th international conference on Human computer interaction with mobile devices and services, pp. 371–374. ACM (2008)
3. Andreopoulos, A., Tsotsos, J.K.: 50 years of object recognition: Directions forward. Computer Vision and Image Understanding **117**(8), 827–891 (2013)

4. Arandjelović, R., Zisserman, A.: Name that sculpture. In: Proceedings of the 2nd ACM International Conference on Multimedia Retrieval, p. 3. ACM (2012)
5. Bacha, S., Benblidia, N.: Combining context and content for automatic image annotation on mobile phones. In: IT Convergence and Security (ICITCS), 2013 International Conference on, pp. 1–4. IEEE (2013)
6. Bay, H., Tuytelaars, T., Van Gool, L.: Surf: Speeded up robust features. In: Computer vision–ECCV 2006, pp. 404–417. Springer (2006)
7. Chandrasekhar, V., Takacs, G., Chen, D., Tsai, S., Grzeszczuk, R., Girod, B.: Chog: Compressed histogram of gradients a low bit-rate feature descriptor. In: Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, pp. 2504–2511. IEEE (2009)
8. Chen, D.M., Baatz, G., Köser, K., Tsai, S.S., Vedantham, R., Pylvä, T., Roimela, K., Chen, X., Bach, J., Pollefeys, M., et al.: City-scale landmark identification on mobile devices. In: Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on, pp. 737–744. IEEE (2011)
9. Chen, D.M., Makar, M., Araujo, A.F., Girod, B.: Interframe coding of global image signatures for mobile augmented reality. In: Data Compression Conference (DCC), 2014, pp. 33–42. IEEE (2014)
10. Chen, D.M., Tsai, S.S., Chandrasekhar, V., Takacs, G., Singh, J., Girod, B.: Tree histogram coding for mobile image matching. In: Data Compression Conference, 2009. DCC'09., pp. 143–152. IEEE (2009)
11. Chen, D.M., Tsai, S.S., Vedantham, R., Grzeszczuk, R., Girod, B.: Streaming mobile augmented reality on mobile phones. In: Mixed and Augmented Reality, 2009. ISMAR 2009. 8th IEEE International Symposium on, pp. 181–182. IEEE (2009)
12. Chen, T., Fan, J., Lu, S.: Context-aware codebook learning for mobile landmark recognition. In: Image Processing (ICIP), 2014 IEEE International Conference on, pp. 3963–3967. IEEE (2014)
13. Chen, T., Li, Z., Yap, K.H., Wu, K., Chau, L.P.: A multi-scale learning approach for landmark recognition using mobile devices. In: Information, Communications and Signal Processing, 2009. ICICS 2009. 7th International Conference on, pp. 1–4. IEEE (2009)
14. Chen, T., Lu, S., Fan, J.: Context-aware vocabulary tree for mobile landmark recognition. Journal of Visual Communication and Image Representation (2015)
15. Chen, T., Yap, K.H.: Context-aware discriminative vocabulary learning for mobile landmark recognition. Circuits and Systems for Video Technology, IEEE Transactions on **23**(9), 1611–1621 (2013)
16. Chen, T., Yap, K.H.: Discriminative bow framework for mobile landmark recognition. Cybernetics, IEEE Transactions on **44**(5), 695–706 (2014)
17. Chen, T., Yap, K.H., Chau, L.P.: Content and context information fusion for mobile landmark recognition. In: Information, Communications and Signal Processing (ICICS) 2011 8th International Conference on, pp. 1–4. IEEE (2011)
18. Chen, T., Yap, K.H., Chau, L.P.: Integrated content and context analysis for mobile landmark recognition. Circuits and Systems for Video Technology, IEEE Transactions on **21**(10), 1476–1486 (2011)
19. Chen, T., Yap, K.H., Zhang, D.: Discriminative soft bag-of-visual phrase for mobile landmark recognition. Multimedia, IEEE Transactions on **16**(3), 612–622 (2014)
20. Chen, W.C., Xiong, Y., Gao, J., Gelfand, N., Grzeszczuk, R.: Efficient extraction of robust image features on mobile devices. In: Proceedings of the 2007 6th IEEE and ACM International Symposium on Mixed and Augmented Reality, pp. 1–2. IEEE Computer Society (2007)
21. Cushen, G., Nixon, M.S., et al.: Mobile visual clothing search. In: Multimedia and Expo Workshops (ICMEW), 2013 IEEE International Conference on, pp. 1–6. IEEE (2013)
22. Di, W., Wah, C., Bhardwaj, A., Piramuthu, R., Sundaresan, N.: Style finder: Fine-grained clothing style detection and retrieval. In: Computer Vision and Pattern Recognition Workshops (CVPRW), 2013 IEEE Conference on, pp. 8–13. IEEE (2013)
23. Duan, L.Y., Ji, R., Chen, J., Yao, H., Huang, T., Gao, W.: Learning from mobile contexts to minimize the mobile location search latency. Signal Processing: Image Communication **28**(4), 368–385 (2013)
24. Duan, L.Y., Ji, R., Chen, Z., Huang, T., Gao, W.: Towards mobile document image retrieval for digital library. Multimedia, IEEE Transactions on **16**(2), 346–359 (2014)
25. Fritz, G., Seifert, C., Paletta, L.: A mobile vision system for urban detection with informative local descriptors. In: Computer Vision Systems, 2006 ICVS'06. IEEE International Conference on, pp. 30–30. IEEE (2006)

26. Girod, B., Chandrasekhar, V., Chen, D.M., Cheung, N.M., Grzeszczuk, R., Reznik, Y., Takacs, G., Tsai, S.S., Vedantham, R.: Mobile visual search. Signal Processing Magazine, IEEE **28**(4), 61–76 (2011)
27. Guan, T., He, Y., Duan, L., Yang, J., Gao, J., Yu, J.: Efficient bof generation and compression for on-device mobile visual location recognition. MultiMedia, IEEE **21**(2), 32–41 (2014)
28. Guan, T., He, Y., Gao, J., Yang, J., Yu, J.: On-device mobile visual location recognition by integrating vision and inertial sensors. Multimedia, IEEE Transactions on **15**(7), 1688–1699 (2013)
29. Gui, Z., Wang, Y., Liu, Y., Chen, J.: Mobile visual recognition on smartphones. Journal of Sensors **2013** (2013)
30. Hao, J., Wang, G., Seo, B., Zimmermann, R.: Point of interest detection and visual distance estimation for sensor-rich video. Multimedia, IEEE Transactions on **16**(7), 1929–1941 (2014)
31. He, J., Feng, J., Liu, X., Cheng, T., Lin, T.H., Chung, H., Chang, S.F.: Mobile product search with bag of hash bits and boundary reranking. In: Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, pp. 3005–3012. IEEE (2012)
32. He, J., Lin, T.H., Feng, J., Chang, S.F.: Mobile product search with bag of hash bits. In: Proceedings of the 19th ACM international conference on Multimedia, pp. 839–840. ACM (2011)
33. Herranz, L., Xu, R., Jiang, S.: A probabilistic model for food image recognition in restaurants. In: Proceedings of the IEEE ICME (2015)
34. Houle, M.E., Oria, V., Satoh, S., Sun, J.: Annotation propagation in image databases using similarity graphs. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM) **10**(1), 7 (2013)
35. Huang, K., Ding, X., Chen, G., Saenko, K.: Automatic mobile photo tagging using context. In: TENCON 2013-2013 IEEE Region 10 Conference (31194), pp. 1–5. IEEE (2013)
36. Ivanov, I., Vajda, P., Goldmann, L., Lee, J.S., Ebrahimi, T.: Object-based tag propagation for semi-automatic annotation of images. In: Proceedings of the international conference on Multimedia information retrieval, pp. 497–506. ACM (2010)
37. Je, S.k., Lee, S., Oh, W.G.: Mobile visual search applications. In: Proceedings of the International Conference on Image Processing, Computer Vision, and Pattern Recognition (IPCV), p. 1. The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp) (2014)
38. Ji, R., Duan, L.Y., Chen, J., Yao, H., Gao, W.: When codeword frequency meets geographical location. In: Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on, pp. 2400–2403. IEEE (2011)
39. Ji, R., Duan, L.Y., Chen, J., Yao, H., Huang, T., Gao, W.: Learning compact visual descriptor for low bit rate mobile landmark search. In: IJCAI Proceedings-International Joint Conference on Artificial Intelligence, vol. 22, p. 2456 (2011)
40. Ji, R., Duan, L.Y., Chen, J., Yao, H., Rui, Y., Chang, S.F., Gao, W.: Towards low bit rate mobile visual search with multiple-channel coding. In: Proceedings of the 19th ACM international conference on Multimedia, pp. 573–582. ACM (2011)
41. Ji, R., Duan, L.Y., Chen, J., Yao, H., Yuan, J., Rui, Y., Gao, W.: Location discriminative vocabulary coding for mobile landmark search. International Journal of Computer Vision **96**(3), 290–314 (2012)
42. Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T.: Caffe: Convolutional architecture for fast feature embedding. In: Proceedings of the ACM International Conference on Multimedia, pp. 675–678. ACM (2014)
43. Kawano, Y., Yanai, K.: Real-time mobile food recognition system. In: Computer Vision and Pattern Recognition Workshops (CVPRW), 2013 IEEE Conference on, pp. 1–7. IEEE (2013)
44. Kawano, Y., Yanai, K.: Foodcam-256: A large-scale real-time mobile food recognitionsystem employing high-dimensional features and compression of classifier weights. In: Proceedings of the ACM International Conference on Multimedia, pp. 761–762. ACM (2014)
45. Kawano, Y., Yanai, K.: Foodcam: A real-time mobile food recognition system employing fisher vector. In: MultiMedia Modeling, pp. 369–373. Springer (2014)
46. Kim, D., Hwang, E., Rho, S.: Location-based large-scale landmark image recognition scheme for mobile devices. In: Mobile, Ubiquitous, and Intelligent Computing (MUSIC), 2012 Third FTRA International Conference on, pp. 47–52 (2012)

47. Kuo, Y.H., Lee, W.Y., Hsu, W.H., Cheng, W.H.: Augmenting mobile city-view image retrieval with context-rich user-contributed photos. In: Proceedings of the 19th ACM international conference on Multimedia, pp. 687–690. ACM (2011)
48. Lee, Y.H., Kim, B., Kim, H.J.: Photograph indexing and retrieval using combined geo-information and visual features. In: Complex, Intelligent and Software Intensive Systems (CISIS), 2010 International Conference on, pp. 790–793. IEEE (2010)
49. Li, Y., Lim, J.H.: Outdoor place recognition using compact local descriptors and multiple queries with user verification. In: Proceedings of the 15th international conference on Multimedia, pp. 549–552. ACM (2007)
50. Li, Z., Yap, K.H.: Content and context boosting for mobile landmark recognition. Signal Processing Letters, IEEE **19**(8), 459–462 (2012)
51. Li, Z., Yap, K.H.: Context-aware discriminative vocabulary tree learning for mobile landmark recognition. Digital Signal Processing **24**, 124–134 (2014)
52. Li, Z., Yap, K.H., Tan, K.W.: Context-aware mobile image annotation for media search and sharing. Signal Processing: Image Communication **28**(6), 624–641 (2013)
53. Lim, J.H., Li, Y., You, Y., Chevallet, J.P.: Scene recognition with camera phones for tourist information access. In: Multimedia and Expo, 2007 IEEE International Conference on, pp. 100–103. IEEE (2007)
54. Lin, J., Wu, V.: Tagging content with metadata pre-filtered by context (2013). URL https://www.google.com/patents/US8370358. US Patent 8,370,358
55. Liu, H., Li, H., Mei, T., Luo, J.: Accurate sensing of scene geo-context via mobile visual localization. Multimedia Systems **21**(3), 255–265 (2015)
56. Liu, H., Mei, T., Li, H., Luo, J., Li, S.: Robust and accurate mobile visual localization and its applications. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM) **9**(1s), 51 (2013)
57. Liu, H., Mei, T., Luo, J., Li, H., Li, S.: Finding perfect rendezvous on the go: accurate mobile visual localization and its applications to routing. In: Proceedings of the 20th ACM international conference on Multimedia, pp. 9–18. ACM (2012)
58. Liu, S., Song, Z., Liu, G., Xu, C., Lu, H., Yan, S.: Street-to-shop: Cross-scenario clothing retrieval via parts alignment and auxiliary set. In: Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, pp. 3330–3337. IEEE (2012)
59. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International journal of computer vision **60**(2), 91–110 (2004)
60. Mai, W., Dodds, G., Tweed, C.: A pda-based system for recognizing buildings from user-supplied images. In: Mobile and Ubiquitous Information Access, pp. 143–157. Springer (2004)
61. Maruyama, T., Kawano, Y., Yanai, K.: Real-time mobile recipe recommendation system using food ingredient recognition. In: Proceedings of the 2nd ACM international workshop on Interactive multimedia on mobile and portable devices, pp. 27–34. ACM (2012)
62. Min, W., Xu, C., Xu, M., Xiao, X., Bao, B.K.: Mobile landmark search with 3d models. Multimedia, IEEE Transactions on **16**(3), 623–636 (2014)
63. Mouine, S., Yahiaoui, I., Verroust-Blondet, A., Joyeux, L., Selmi, S., Goëau, H.: An android application for leaf-based plant identification. In: Proceedings of the 3rd ACM conference on International conference on multimedia retrieval, pp. 309–310. ACM (2013)
64. Naaman, M., Nair, R.: Zonetag's collaborative tag suggestions: What is this person doing in my phone? MultiMedia, IEEE **15**(3), 34–40 (2008)
65. Naaman, M., Paepcke, A., Garcia-Molina, H.: From where to what: Metadata sharing for digital photographs with geographic coordinates. In: On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE, pp. 196–217. Springer (2003)
66. O'Hare, N., Gurrin, C., Jones, G.J., Smeaton, A.F.: Combination of content analysis and context features for digital photograph retrieval (2005)
67. Oliveira, L., Costa, V., Neves, G., Oliveira, T., Jorge, E., Lizarraga, M.: A mobile, lightweight, poll-based food identification system. Pattern Recognition **47**(5), 1941–1952 (2014)
68. Panda, J., Sharma, S., Jawahar, C.: Heritage app: annotating images on mobile phones. In: Proceedings of the Eighth Indian Conference on Computer Vision, Graphics and Image Processing, p. 3. ACM (2012)
69. Pei, D., Ji, R., Sun, F., Liu, H.: Estimating viewing angles in mobile street view search. In: Image Processing (ICIP), 2012 19th IEEE International Conference on, pp. 441–444. IEEE (2012)

70. Proß, B., Schöning, J., Krüger, A.: ipiccer: automatically retrieving and inferring tagged location information from web repositories. In: Proceedings of the 11th International Conference on Human-Computer Interaction with Mobile Devices and Services, p. 69. ACM (2009)
71. Qin, C., Bao, X., Choudhury, R.R., Nelakuditi, S.: Tagsense: Leveraging smartphones for automatic image tagging. Mobile Computing, IEEE Transactions on **13**(1), 61–74 (2014)
72. Quack, T., Bay, H., Van Gool, L.: Object recognition for the internet of things. In: The Internet of Things, pp. 230–246. Springer (2008)
73. Ruf, B., Kokiopoulou, E., Detyniecki, M.: Mobile museum guide based on fast sift recognition. In: Adaptive Multimedia Retrieval. Identifying, Summarizing, and Recommending Image and Music, pp. 170–183. Springer (2010)
74. Runge, N., Wenig, D., Malaka, R.: Keep an eye on your photos: automatic image tagging on mobile devices. In: Proceedings of the 16th international conference on Human-computer interaction with mobile devices & services, pp. 513–518. ACM (2014)
75. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision (IJCV) pp. 1–42 (2015). DOI 10.1007/s11263-015-0816-y
76. Sang, J., Mei, T., Xu, Y.Q., Zhao, C., Xu, C., Li, S.: Interaction design for mobile visual search. Multimedia, IEEE Transactions on **15**(7), 1665–1676 (2013)
77. Schroth, G., Huitl, R., Abu-Alqumsan, M., Schweiger, F., Steinbach, E.: Exploiting prior knowledge in mobile visual location recognition. In: Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on, pp. 2357–2360. IEEE (2012)
78. Schroth, G., Huitl, R., Chen, D., Abu-Alqumsan, M., Al-Nuaimi, A., Steinbach, E.: Mobile visual location recognition. Signal Processing Magazine, IEEE **28**(4), 77–89 (2011)
79. Seifert, C., Paletta, L., Jeitler, A., Hödl, E., Andreu, J.P., Luley, P., Almer, A.: Visual object detection for mobile road sign inventory. In: Mobile Human-Computer Interaction-MobileHCI 2004, pp. 491–495. Springer (2004)
80. Shen, X., Lin, Z., Brandt, J., Wu, Y.: Mobile product image search by automatic query object extraction. In: Computer Vision–ECCV 2012, pp. 114–127. Springer (2012)
81. Sinha, P., Jain, R.: Classification and annotation of digital photos using optical context data. In: Proceedings of the 2008 international conference on Content-based image and video retrieval, pp. 309–318. ACM (2008)
82. Song, X., Jiang, S., Xu, R., Herranz, L.: Semantic features for food image recognition with geo-constraints. In: Data Mining Workshop (ICDMW), 2014 IEEE International Conference on, pp. 1020–1025. IEEE (2014)
83. Takacs, G., Chandrasekhar, V., Gelfand, N., Xiong, Y., Chen, W.C., Bismpigiannis, T., Grzeszczuk, R., Pulli, K., Girod, B.: Outdoors augmented reality on mobile phone using loxel-based visual feature organization. In: Proceedings of the 1st ACM international conference on Multimedia information retrieval, pp. 427–434. ACM (2008)
84. Tsai, C.M., Qamra, A., Chang, E.Y., Wang, Y.F.: Extent: Inferring image metadata from context and content. In: Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on, pp. 1270–1273. IEEE (2005)
85. Tsai, S.S., Chen, D., Chandrasekhar, V., Takacs, G., Cheung, N.M., Vedantham, R., Grzeszczuk, R., Girod, B.: Mobile product recognition. In: Proceedings of the international conference on Multimedia, pp. 1587–1590. ACM (2010)
86. Tsai, S.S., Chen, D., Takacs, G., Chandrasekhar, V., Singh, J.P., Girod, B.: Location coding for mobile image retrieval. In: Proceedings of the 5th International ICST Mobile Multimedia Communications Conference, p. 8. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering) (2009)
87. Viana, W., Braga, R., Lemos, F.D., de Souza, J.M., Carmo, R., Andrade, R., Martin, H., et al.: Mobile photo recommendation and logbook generation using context-tagged images. MultiMedia, IEEE **21**(1), 24–34 (2014)
88. Xia, J., Gao, K., Zhang, D., Mao, Z.: Geometric context-preserving progressive transmission in mobile visual search. In: Proceedings of the 20th ACM international conference on Multimedia, pp. 953–956. ACM (2012)
89. Xiao, J., Hays, J., Ehinger, K., Oliva, A., Torralba, A., et al.: Sun database: Large-scale scene recognition from abbey to zoo. In: Computer vision and pattern recognition (CVPR), 2010 IEEE conference on, pp. 3485–3492. IEEE (2010)
90. Xie, X., Lu, L., Jia, M., Li, H., Seide, F., Ma, W.Y.: Mobile search with multimodal queries. Proceedings of the IEEE **96**(4), 589–601 (2008)

91. Xu, R., Herranz, L., Jiang, S., Wang, S., Song, X., Jain, R.: Geolocalized modeling for dish recognition. Multimedia **In press**, IEEE Transactions on
92. Yang, D.S., Lee, Y.H.: Mobile image retrieval using integration of geo-sensing and visual descriptor. In: Network-Based Information Systems (NBiS), 2012 15th International Conference on, pp. 743–748. IEEE (2012)
93. Yap, K.H., Chen, T., Li, Z., Wu, K.: A comparative study of mobile-based landmark recognition techniques. Intelligent Systems, IEEE **25**(1), 48–57 (2010)
94. You, Q., Yuan, J., Wang, J., Guo, P., Luo, J.: Snap n'shop: Visual search-based mobile shopping made a breeze by machine and crowd intelligence. In: Semantic Computing (ICSC), 2015 IEEE International Conference on, pp. 173–180. IEEE (2015)
95. Yu, F.X.: Intelligent query formulation for mobile visual search. In: Proceedings of the 19th ACM international conference on Multimedia, pp. 861–862. ACM (2011)
96. Yu, F.X., Ji, R., Chang, S.F.: Active query sensing for mobile location search. In: Proceedings of the 19th ACM international conference on Multimedia, pp. 3–12. ACM (2011)
97. Zamir, A.R., Dehghan, A., Shah, M.: Visual business recognition: a multimodal approach. In: ACM Multimedia, pp. 665–668. Citeseer (2013)
98. Zhang, C., Zhang, Y., Zhu, X., Xue, Z., Qin, L., Huang, Q., Tian, Q.: Socio-mobile landmark recognition using local features with adaptive region selection. Neurocomputing (2015)
99. Zhang, N., Mei, T., Hua, X.S., Guan, L., Li, S.: Interactive mobile visual search for social activities completion using query image contextual model. In: Multimedia Signal Processing (MMSP), 2012 IEEE 14th International Workshop on, pp. 238–243. IEEE (2012)
100. Zhu, C., Li, K., Lv, Q., Shang, L., Dick, R.P.: iscope: personalized multi-modality image search for mobile devices. In: Proceedings of the 7th international conference on Mobile systems, applications, and services, pp. 277–290. ACM (2009)