Multimedia Systems (2017) 23:647–665
DOI 10.1007/s00530-016-0523-8

SPECIAL ISSUE PAPER

A survey on context-aware mobile visual recognition

Weiqing Min[1] · Shuqiang Jiang[1] · Shuhui Wang[1] · Ruihan Xu[1] · Yushan Cao[2] · Luis Herranz[1] · Zhiqiang He[3]

Published online: 7 July 2016

Abstract The phenomenal growth of the usage of mobile devices (e.g., mobile phones and tablet PCs) opens up a new service, namely mobile visual recognition, which has been widely used in many areas, such as mobile shopping and augmented reality. The rich contextual 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 Context-Aware Mobile Visual Recognition (CAMVR) and reviews related work regarding to different contextual information, recognition methods, recognition types, and various application scenarios. Finally, we discuss 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 to as mobile visual recognition. For example, a user takes a photo of a landmark and automatically obtains the textual information (e.g., landmark tags and relevant descriptions), related images (e.g., different views of the same landmark), or a 3D model [73] about the landmark. Mobile visual recognition is particularly useful in applications, such as mobile shopping [40, 68], mobile landmark recognition for tourists [11], and mobile location recognition for augmented reality [94]. Furthermore, such mobile visual recognition functionalities have been shown in many commercial systems, such as Google "Goggles",[1] Amazon "Snaptell",[2] and "Kooaba".[3]

Because of its great potential in the industry, mobile visual recognition has received increasing attention in academia. Girod et al. [33] proposed a complete mobile visual search system, including feature extraction, feature matching, and geometry verification. For each block of the search pipeline, they designed their solutions different

✉ Shuqiang Jiang
sqjiang@ict.ac.cn

Weiqing Min
weiqing.min@vipl.ict.ac.cn

Shuhui Wang
wangshuhui@ict.ac.cn

Ruihan Xu
rhxu@ict.ac.cn

Yushan Cao
caoyushan@enet.edu.cn

Luis Herranz
luis.herranz@vipl.ict.ac.cn

Zhiqiang He
lirong2@lenovo.com

[1] Key Lab of Intelligent Information Processing, Institute of Computing Technology, CAS, Beijing 100190, China

[2] Higher Education Institution Teacher Online Training Center, Beijing, China

[3] Lenovo Corporate Research, Beijing 100085, China

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

from general visual recognition to facilitate mobile visual search. Furthermore, they released a data set for performance evaluation. Chatzilari et al. [10] performed an extensive comparative study of different recognition approaches on the mobile device by evaluating the performance of the feature extraction and encoding algorithms. Compared with general visual recognition, mobile visual recognition has its unique challenges:

- *Limited network bandwidth* With the development of the Internet communicate technology, such as 4G, the bandwidth of networks increased fast. However, there is still a bottleneck in many areas, especially those densely populated ones, where many people are using mobile devices simultaneously. Many mobile visual systems extract features in the mobile side. However, the amount of visual features sent from the mobile side to the server should be reduced to satisfy the real-time query requirement, which probably leads to the degradation of the recognition performance. Therefore, under the limitation of the network bandwidth, how to send compressed features without affecting the recognition performance is a challenging problem in the mobile recognition environment.
- *Limited battery power* Existing mobile devices have limited capacity of the power. Sending a feature vector of the query image saves network bandwidth and further reduces the transmission cost. However, computing features will consume the power of the battery significantly. Obviously, this challenges the tolerant attitudes of users to a short battery running time, since recharging is usually inconvenient for users, especially when they are traveling.
- *Diverse photo-taking conditions* Because of different camera configurations in the mobile device (e.g., different resolutions) and diverse indoor/outdoor conditions (e.g., varying weather conditions), how to achieve robust visual recognition under these conditions is also very challenging.

To solve these problems, many existing works [13, 33, 41, 108] have developed different visual recognition methods to improve the mobile visual recognition experience. These methods directly extract the visual features for image representation, including deep features [52]. To reduce the amount of data sent from the mobile device to the server, some encoding methods on the mobile side have been developed to compress the visual features, such as SURF [7], CHoG [8], and BoHB [40]. However, one shortcoming of these approaches is that they mainly analyze the content alone, while ignore the rich contextual information (e.g., the GPS and time information) easily acquired by the mobile device, which can speed up the recognition time and improve the recognition performance.

In fact, mobile devices bring a lot of contextual information, which can be categorized into two levels: one is the internal contextual 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 contextual information which could be easily acquired by the mobile device, such as time and geo-location. Researchers have exploited many of them to improve the recognition performance. Commonly used contexts include location, direction, time, text, gravity, acceleration, and other camera parameters. For example, in [95], content analysis is essentially filtered by a pre-defined area centered at the GPS location of the query image. Chen et al. [11] utilized the GPS information to narrow the search space for landmark recognition. Ji et al. [51] 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, 22] combined the visual information with the contextual information, including the location and the direction information for mobile landmark recognition. Runge et al. [86] suggested the tags of images using the location name and time period. Gui et al. [36] fused outputs of inertial sensors and computer vision techniques for mobile scene recognition. In such cases, utilizing the contextual information in mobile visual recognition can speed up the recognition time and improve the recognition performance.

In this survey, we give a comprehensive overview of Context-Aware Mobile Visual Recognition (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 contextual 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 contextual information, recognition method, and 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 Sect. 2 through Sect. 4, we survey the state-of-the-art approaches of CAMVR according to different contextual information, different recognition methods, and different recognition types, respectively. In Sect. 5, we introduce various application scenarios based on CAMVR. Finally, we conclude the paper with a discussion of future research directions in Sect. 6.

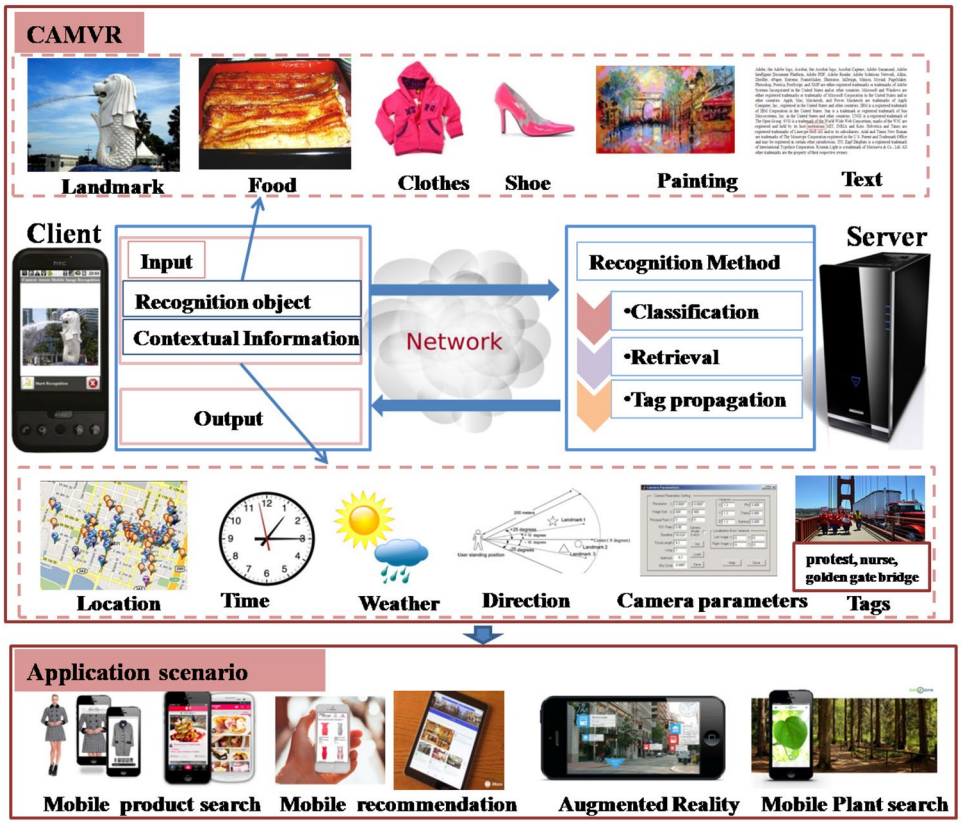

Fig. 1 Overview of CAMVR. *Top* the flowchart of CAMVR; *bottom* application scenarios of CAMVR

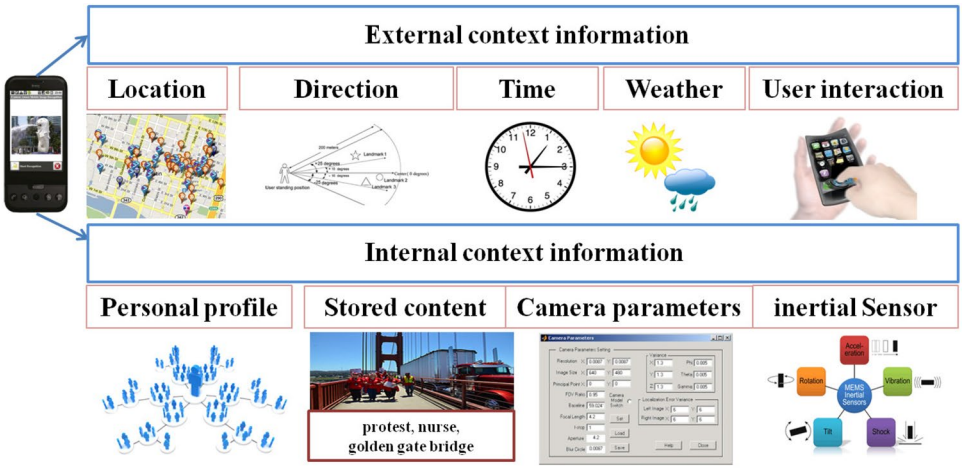

Fig. 2 Different kinds of contextual information

2 Types of contextual information

In this section, we review related work on CAMVR based on different types of contextual information. As shown in Fig. 2, the contextual 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, stored textual/visual content, and camera's parameters. CAMVR can exploit various forms of contextual 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, greatly improves recognition accuracy and speed [11]. Direction refers to the shooting direction, a necessary complement for location information, especially for recognizing distant target or scene [22]. As lots of images, especially those uploaded to social networks, contain text descriptions or tags input by

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

Contextual information	Location	Context with location	Context without location
Representative work	Tsai et al. [95]		
	Fritz et al. [32]	Benjamin et al. [82]	Li et al. [59]
	Quack et al. [84]	Naaman et al. [75]	Xia et al. [100]
	Takacs et al. [95]	Sinha et al. [93]	Zhang et al. [112]
	Zhu et al. [112]	Chen et al. [11]	Gui et al. [36]
	Yap et al. [104]	Li et al. [64]	Hao et al. [38]
	Chen et al. [21]	Chen et al. [18, 22]	Qin et al. [83]
	Ji et al. [49, 51]	Guan et al. [34, 35]	You et al. [106]
	Duan et al. [30]		

users, text contexts play an important role in recognizing uploaded images [1].

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

- *Location* Only the location information is used to improve the recognition [57, 109].
- *Context with location* Besides the location information, other contextual information (e.g., time and direction) is also employed, which are probably complimentary for location information [18, 19].
- *Context without location* This type of contextual information, such as inertial sensors' parameters, is used for CAMVR [36].

2.1 Location

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, can speed up the recognition time and improve the recognition accuracy [104]. For example, Takacs et al. [94] 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 [2, 94], Kuo et al. [57] also introduced the GPS constraints in the retrieval process on inverted indexing, so that they can satisfy the requirement of a real-time image retrieval system. In [32, 78, 95], 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 areas can be reduced substantially. Xie et al. [101] proposed a multi-modal search scheme which uses the image content and user location to increase the search accuracy, while Zhu et al. [112] 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. [109] proposed a multi-modal 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, Maiet al. [70] 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 [11, 30, 42, 48, 49, 51, 56, 84, 93, 102, 104, 110] focuses on utilizing the GPS information for specific tasks, such as roadside sign recognition [90], mobile landmark recognition [11, 30, 48, 49, 51, 56, 61, 84, 104], and mobile food recognition [42, 93, 102]. For example, Seifertet al. [90] proposed a mobile system based on a GPS sensor for roadside sign localization and classification. Chen et al. [11] utilized the GPS coordinates to narrow the search space for landmark recognition. Jiet al. [48, 49, 51] designed a GPS-based location discriminative vocabulary coding scheme, which achieves extremely low-bit-rate query transmission for mobile landmark search. Song et al. [93] introduced geo-constraints for food image recognition.

However, GPS-based mobile visual recognition has some drawbacks [11, 67, 89] that make it impractical in real applications: first, the embedded GPS modules rely on a satellite navigation system and need at least four satellites to provide sufficient positioning 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 m. The large GPS error of the captured image will result in wrong recognition. Second, besides the GPS information, there are other contextual information available

from the mobile devices. The effective integration of different contextual information will further improve the recognition performance. Therefore, some work [11, 67] has resorted to combining the GPS information with other contextual information (e.g., direction information) to enhance the recognition performance.

2.2 Context with location

In addition to the GPS information, other contextual information, such as direction and time information, can be easily acquired by mobile devices equipped with digital compass and other sensors. Combining the content information with richer contextual information will improve the recognition performance. For example, Benjamin et al. [82] presented a system iPiccer to infer photo tags from its location and orientation. Chen et al. [15, 17–22] incorporated the location and direction information to perform mobile landmark recognition. Direction information is obtained through the built-in digital compass of mobile devices and is complementary to the location information. Similarly, Li et al. [60] proposed a boosting algorithm to integrate visual content and two types of contextual information, including the location and direction for mobile landmark recognition. Guanet al. [34, 35] 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. [58, 103] utilized the geographic location and time where the photo was taken to create automatic spatial and temporal indexes for image retrieval. Lin et al. [64] generated tags for content from meta-data, which is pre-filtered based on the location and time information. Runge et al. [86] proposed a method to use the location name and time period to suggest tags of images.

Furthermore, the integration of more than two kinds of contextual information with location information is also utilized for mobile visual recognition. Ahern et al. [1] proposed a media annotation system via various contextual information, including restaurants, events, venues near the user's location, past tags from the user, and the user's social network. Naaman et al. [75] balanced all the tag sources to generate a prioritized suggested tag list using several contextual information, including the location, the tags' social context, and temporal context. Li et al. [62] utilized three types of contextual information, namely location, user interaction, and Web for mobile image annotation. Huang et al. [44] utilized clustering and similarity-based approaches for photo tagging using various contextual information, such as date, time, location, environment noise, and human faces. Pinaki et al. [92] conducted photo annotation by exploiting the following four kinds of meta-information: optical meta-layer, which contains the meta-data related to the optics of the camera, e.g., the focal length and exposure time; temporal meta-layer, which contains the time stamp of the instant where the photo was taken; spatial meta-layer, which contains the spatial coordinates of the places where pictures were shot; and human induced meta-layer, which contains the tags and comments posted by people.

2.3 Context without location

Some specific mobile applications do not need location information but other contextual information, such as camera and other sensors' parameters. For example, Gui et al. [36] fused outputs of inertial sensors and computer vision techniques for mobile scene recognition. Hao et al. [38] 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. [81] studied viewing angle estimation by exploiting the visual appearance of the query, which can be further improved by incorporating coarse mobile context, such as gyro or compass information. Xia et al. [100] proposed an effective and efficient geometric context-preserving progressive transmission method for mobile visual search. Qin et al. [83] 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 [59, 87, 106, 111] considered the user interaction as the contextual information for mobile image retrieval.

3 Recognition methods

Existing recognition methods of CAMVR can be summarized into the following three categories, namely classification-based methods, retrieval-based methods, and tag propagation-based methods. Table 2 summarizes representative work for each category.

3.1 Classification-based methods

Classification-based methods first train a recognizer for each object (e.g., landmark and food) by integrating content and context analysis, and then recognizes the query image using the trained classifier and the contextual 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), contextual information extraction (e.g., determining the location through GPS), and classification (identifying which category the captured object belongs to). In the

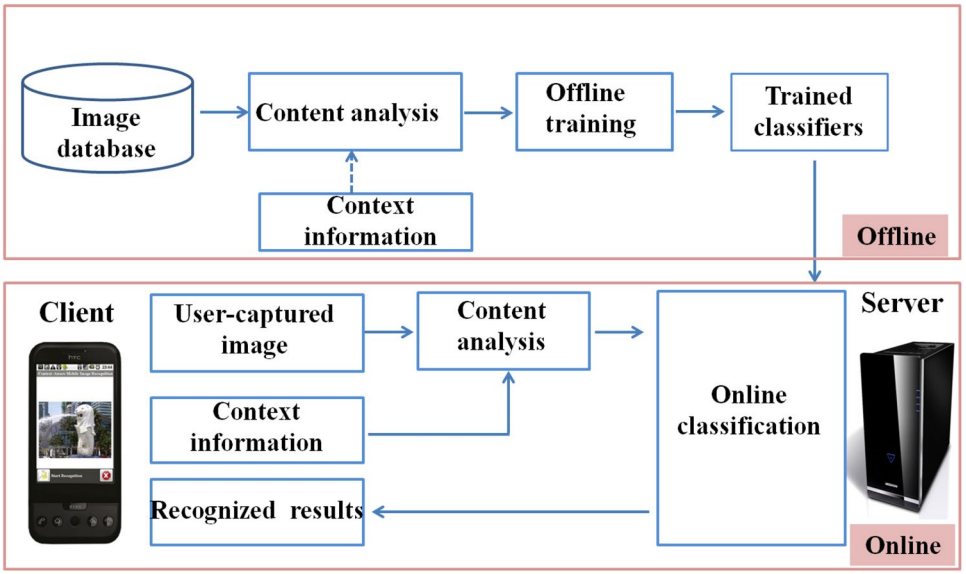

Fig. 3 Mobile visual classification system overview

Table 2 Summarization of CAMVR based on different recognition methods

Recognition method	Classification	Retrieval	Tag propagation
Representative work	Fritz et al. [32]		
	Lim et al. [63]		
	Chen et al. [16]	Girod et al. [33]	Naaman et al. [76]
	Chen et al. [20]	Yu et al. [108]	Ahern et al. [1]
	Li et al. [60]	Chenet al. [22]	Arandjelović et al. [3]
	Chen et al. [11]	He et al. [40]	Li et al. [62]
	Xu et al. [102]		

following sections, we present the state-of-the-art content analysis with contextual information and classification algorithms, respectively.

Content analysis mainly extracts features from the image. We can broadly categorize the visual features into global and local features [104]. 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 [19] can be divided into two classes: (1) local patch image representation [16, 63] that uses visual features extracted from the local patches in the image for recognition and (2) bag-of-words (BoW) histogram representation [21, 32, 40, 97, 108] that generates a BoW histogram for each image through vector quantization.

For local patch representation, Lim et al. [63] employed a discriminative patch selection algorithm to extract 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. To solve this problem, Chen et al. [16] first 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 the 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 200–600 dimensions [104]. Thus, BoW is more suitable for real-time mobile landmark recognition.

The SIFT descriptor [69] is one of the most widely used local descriptors in state-of-the-art visual recognition methods [11, 32, 59, 108]. However, the conventional SIFT involves the detection of a large number of salient keypoints and extraction of a 128-dimensional feature vector centered on each of these keypoints. Because of the limitation of the storage and power in mobile device, the SIFT descriptors are not suitable in the context of mobile visual search applications. To reduce the computational cost of standard SIFTs, researchers have proposed several SIFT-based variants that reduce the number of keypoints and feature dimensions, such as clustering to group similar keypoints [59], Informative-SIFT [32], and SURF [23, 94]. Chandrasekhar et al. [8] further proposed CHoG descriptors with a 20× reduction in bit rate compared to state-of-the art descriptors. In contrast to SIFT and SURF, CHoG coarsely quantizes the 2D gradient histogram and captures the histogram directly into the descriptors with

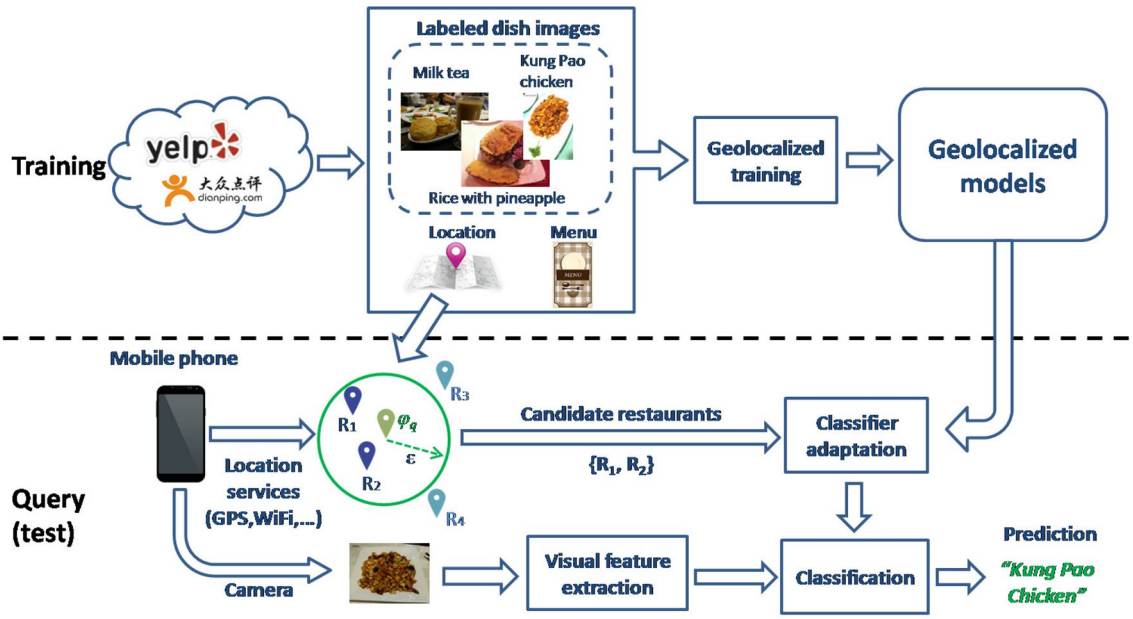

Fig. 4 Overview of the framework for dish recognition with geo-localized models [102]

Huffman and Gagie trees to create very low-bit-rate descriptors. Different from CHoG, He et al. [40] proposed "Bag of Hash Bits" (BoHB), where each local feature is encoded to tens of hash bits using similarity preserving hashing functions, and each image is then represented as a bag of hash bits instead of bag of words. Such BoHB method leverages the distinct properties of hashing bits, such as multi-table indexing, multiple bucket probing, bit reuse, and hamming distance-based ranking to significantly outperform CHoG.

Integrating contextual information into content analysis can make the features more discriminative. Note that the contextual information is not only used in the online phase, but also offline feature learning. For CAMVR, most works [11, 19–22] mainly use the contextual information to reduce the search space for the query image in the online phase. During the offline learning phase, the contextual information is not incorporated. However, some work [17, 18, 51, 60] also fully utilized the contextual information in the offline discriminative feature learning. For example, Ji et al. [51] 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. [60] improved the content-based recognition performance by incorporating recognition results from various context-based vocabulary trees (VTs) built upon location and direction contextual information. Similar to [60], Chen et al. [17, 18] exploited both location and direction information to learn a discriminative compact vocabulary (DCV). Xu et al. [42, 93, 102] combined the GPS information and visual information for discriminative training and food recognition.

For the classification methods, most works [16, 20–22, 63, 104] employs support vector machine (SVM), 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. [21] employed multi-class support vector machine (SVM) with the new spatial pyramid kernel (SPK) to train the landmark classifiers. In [19], they used ensemble of classifiers with fuzzy support vector for training. Xu et al. [102] adopted a location-adaptive SVM classification for training. In contrast, Li et al. [60] 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. Fritz et al. [32] applied MAP classification to mobile visual applications. Yap et al. [104] provided a general overview of existing mobile-based and non-mobile-based landmark recognition systems and their differences. They discussed content and context analyses and compared landmark classification methods. They also presented the experimental results of their own mobile landmark recognition evaluations based on content analysis, context analysis, and integrated content–context analysis.

As representative work, Xu et al. [102] proposed to use restaurants as geographical anchors for mobile dish recognition. As shown in Fig. 4, the method first constructs a

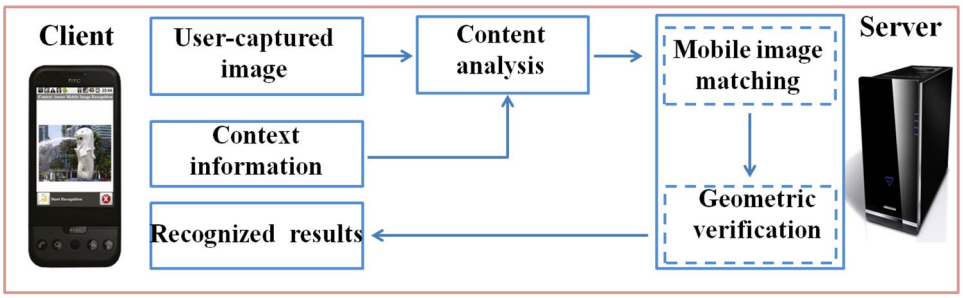

Fig. 5 Mobile visual retrieval system overview

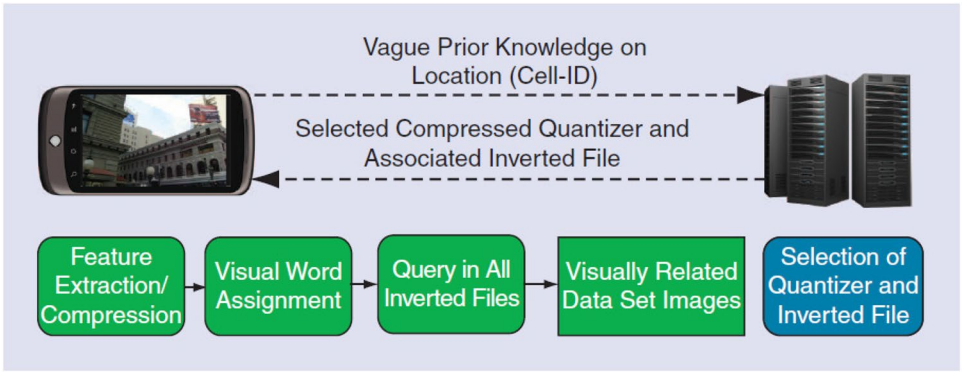

Fig. 6 A pipeline for a visual location recognition system [89]

database of restaurants, including geographical locations and menus, obtained from restaurant review Websites, which also include images of the corresponding dishes. Then, the geo-localized models are trained with these images for each dish, where each model is related to a particular geo-location. During test time, the particular geo-location of the query defines a neighborhood with some candidate restaurants. For each query, the corresponding geo-localized models are selected and combined into a new classifier adapted to the query. They designed two strategies to implement this approach. The first strategy is to train multiple binary pairwise classifiers (also known as one-against-one classifiers). Then, the input feature is classified by all of them based on a simple geo-localized voting method. The second strategy is to train geo-localized one-against-all models. The experimental results verified the advantage of contextual information in improving recognition performance over visual-only methods.

3.2 Retrieval-based methods

Feature representation in retrieval-based methods is similar to classification-based methods. The difference is that classification-based methods train the model on a training set and use the trained model to conduct recognition, while retrieval-based methods return the "closest" images to the input image from database by feature matching algorithm and then assign the closest label to the input image. Figure 5 shows an overview of a retrieval-based CAMVR system, consisting of content analysis, contextual information extraction, image matching, and geometrical verification. Since content analysis and contextual information extraction are similar to the classification-based methods, we mainly review feature matching and geometric verification (GV) algorithm [33, 94]. For general methods, feature matching finds a small set of images in the database that has many 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 [33]. There are also some different image matching strategies.

Girod et al. [33] discussed each block of the retrieval pipeline, including feature extraction, feature matching, and GV in mobile product recognition (e.g., such as books and DVDs). For feature extraction, to obtain low-rate bits descriptors, they designed a new descriptor-CHoG by quantizing and encoding gradient histograms with Huffman and Gagie trees. It achieves 20× reduction in bit rate compared with SIFT and SURF. In feature matching step, to support the popular Vocabulary Tree(VT)-scoring framework [77], they developed an inverted index compression methods for both hard-binned and soft-binned VTs, which can be used to quickly compare images in a large database against a query image. In GV stage, they used the location information of query and database features to confirm that the feature matches are consistent with a change in viewpoint between the two images. To speed up the response time, the authors

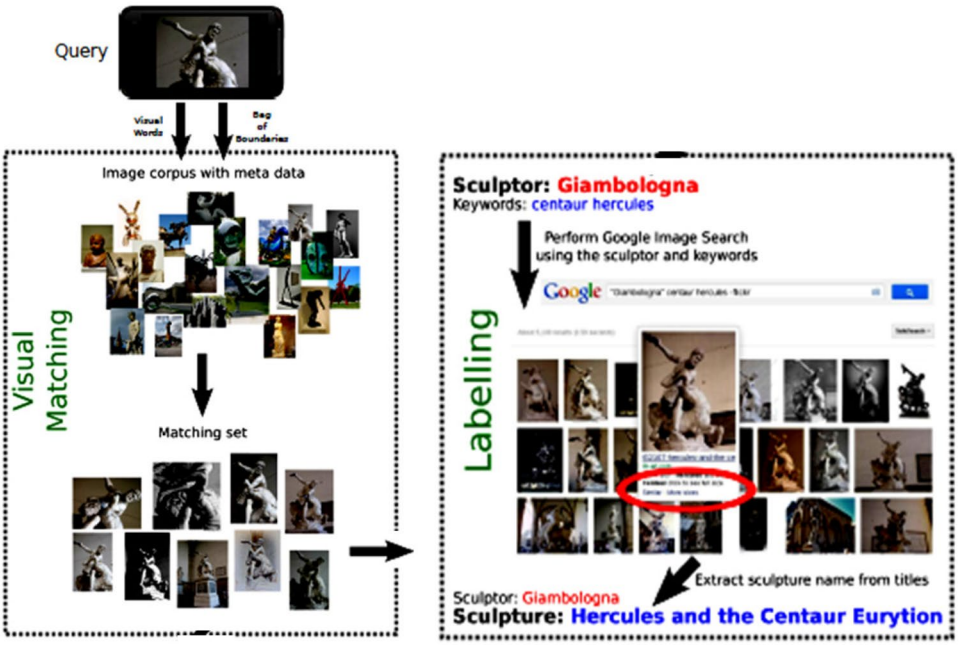

Fig. 7 Overview of mobile sculpture annotation [3]

introduce the prior knowledge of the query location, derived from Cell-IDs. Based on this prior, they proposed a mobile visual location recognition system (Fig. 6) [89]. In this system, they used the Cell-ID of the network provider to determine a position estimate in the range of some hundred meters at most, and then segment the search area into several overlapping subregions for which individual quantization structures and associated inverted file sets are generated. Integrating this prior knowledge into the location recognition process reduces the required resources with respect to memory as well as query time, and increases precision.

As another representative work, Chen et al. [11] proposed a city-scale location recognition system, which improves the mobile landmark identification on mobile devices by fusing two representations of street-level image data, namely perspective central images (PCI) and perspective frontal images (PFI), which contain complementary information. When a query image is taken, it is processed in two parallel pipelines, one for PCIs and the other for PFIs. In the PCI pipeline, the database PCIs are scored using a vocabulary tree trained on SIFT descriptors, geographically distant landmarks are excluded using GPS coordinates associated with the query image (when GPS is available), and geometric verification is performed on a shortlist of database candidates. The PFI pipeline works in a similar way. Another important contribution is that they released a large set of 1.7 million images with GPS information, ground truth labels, and calibration data. The experiments show that the hybrid scheme noticeably boosts recall compared to either PCIs or PFIs by about 10% for both the GPS-aware and GPS-agnostic modes.

In addition, Yu et al. [108] presented a mobile application that can teach mobile users to capture pictures which can distinctively represent the surrounding scenes. Besides feature matching with hash bits and geometry verification, He et al. [40] further introduced the boundary reranking algorithm to improve the retrieval performance. In addition, Arandjelović et al. [4, 27] proposed to improve object retrieval through: (1) replacing the standard Euclidean distance with a square root kernel, (2) discriminative query expansion, and (3) using the spatial verification for feature augmentation.

3.3 Tag propagation-based methods

In contrast to retrieval-based methods, after finding images with tags, which are similar to the query image based on the content or contextual information, tag propagation methods [3, 5, 43, 45] further annotate the query image by propagating the tags of these similar images.

Arandjelović et al. [3] proposed a framework (Fig. 7) to identify sculptures from a query image based on the following two stages: (1) visual matching to a large data set of images of sculptures, and (2) textual labeling given a set of matching images with annotations. In the first stage, they use two complementary visual retrieval methods (one based on visual words and the other on boundaries) to improve both retrieval and precision performance. In the second stage, they proposed a simple voting scheme on the tf-idf weighted meta-data, that can correctly hypothesize a subset of the sculpture name.

In addition, some work resorts to contextual information to restrict the tag propagation process for visual

Table 3 Summarization of CAMVR based on different recognition types

Recognition type	Location	Product	Other objects
Representative work	Schroth et al. [89]		
	Girod et al. [33]		
	Yu et al. [107, 108]		
	Liu et al. [66, 67]	Tsai et al. [96]	
	Duan et al. [30]	He et al. [40, 41]	
	Guan et al. [34]	Shen et al. [91]	Auack et al. [84]
	Liu et al. [65]		Gui et al. [36]
			Mouine et al. [74]
	Chen et al. [11]	Maruyama et al. [71]	Duan et al. [31]
	Ji et al. [51]	Kawano et al. [53]	Xu et al. [102]
	Chen et al.[20, 21]	Liu et al. [68]	
	Chen et al. [18]	Di et al. [29]	
	Chen et al. [19, 22]		
	Min et al. [73]		
	Zhang et al. [110]		

annotation. For example, Naaman et al. [76] 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 contextual information. Li et al. [62] utilized the contextual information, such as the location information, direction information, time information, domain information (e.g., interaction between the user and information server), and Web information to restrict the tag propagation process for image annotation. They also considered different tag distributions at different places in propagating tags to the query images. In addition, Guillaumin et al. [37] proposed a discriminatively trained nearest neighbor model to predict tags by taking a weighted combination of the tag absence/presence among neighbors.

In summary, one key of all three kinds of mobile visual recognition methods with context is how to effectively integrate the contextual information into the content information to reduce the recognition time and improve the recognition performance. However, each type has its application scenario. Classification-based methods aim at determining the class or category of the query, for which a number of training samples are provided and an extra training process is often required. Retrieval-based methods rank a large number of candidates according to their relevance to the query. Tag propagation-based methods return a list of annotated tags by propagating the tags of other images, which are similar to the query image based on the content or contextual information.

4 Recognition types

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

4.1 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. [33, 96] presented a fast and scalable mobile product recognition system, where the database primarily comprises products packaged in rigid boxes with printed labels, such as CDs, DVDs, and books. He et al. [40, 41] encoded each local feature into a very small number of hash bits for efficient mobile product search on different product data sets, which are crawled from online shopping companies, such as Ebay.com, Zappos.com, and Amazon.com. Shen et al. [91] simultaneously retrieved visually similar product images, and localized the product instance in the query image for mobile product images retrieval. Chi et al. [25] proposed a novel representation method, visual part-based object representation for commercial item image recognition and recommendation. Furthermore, Chi et al. [26] developed a mobile-sensing framework for simultaneous object recognition and localization, and have verified its effectiveness in instantly retrieving relevant information of the recognized businesses.

Among all products, mobile food recognition and mobile clothes recognition are particularly useful for great business potentials. For mobile food recognition, Maruyama et al. [71] proposed a system which extracts color features and recognizes 30 kinds of food ingredients on a mobile device. Kawano et al. [53] proposed a real-time food recognition system, where a user first draws bounding boxes by touching the screen, and then the system starts food item recognition within the indicated bounding boxes. Kawano et al. [54, 55] computed Fisher vectors over HOG patches to develop a real-time mobile food recognition system on a larger food data set. Oliveira et al. [79] 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. [102] proposed a framework incorporating discriminative classification in geo-localized settings and introduced the concept of geo-localized models for food recognition. For mobile clothes recognition, Liu et al. [68] proposed a "street-to-shop" clothing retrieval model, where a user takes a photo of any person and retrieve similar clothing from online shops 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, and sweater). Di et al. [29] 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. [28] 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.

4.2 Mobile location recognition

Most interesting location-based services (LBSs) could be provided in densely populated environments, including urban and indoor scenarios [89]. However, GPS is hardly available in these urban streets 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. Girod et al. [33] used compact feature descriptors and spatial coding schemes for mobile visual search, which also proves very useful for vision-based mobile localization. Yu et al. [107, 108] presented a mobile location search application that can teach mobile users to capture pictures that can distinctively represent surrounding scenes. Duan et al. [30] proposed a method to learn an extremely compact visual descriptors from the mobile contexts towards low-bit-rate mobile location search. Tao et al. [34, 35] proposed a memory-and computation-efficient encoding algorithm to enable efficient on-device mobile visual location recognition. Schroth et al. [80, 88] partitioned the whole work space into overlapped subregions and designed a strategy based on prior knowledge (such as Cell-ID) 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. [65–67] proposed a framework 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 attention for its applications in travel recommendation. Chen et al. [11] built up a million-scale street view image data set and conducted concrete experiments to evaluate their landmark retrieval scheme. Ji et al. [50, 51] proposed a discriminative vocabulary coding scheme for mobile landmark search. Similar to [50, 51], Zhang et al. [110] also proposed a method to learn a geo-discriminative codebook for mobile landmark recognition. Besides the location information, Chen et al. [15, 17–22] incorporated the direction information to perform mobile landmark recognition. Similarly, Li et al. [60] used these two types of mobile context: location and direction information for mobile landmark recognition. Different from these work, Min et al. [73] 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 and ends with a 3D model search result.

4.3 Other mobile object recognition

To integrate mobile visual search techniques into a digital library, Duan et al. [31] proposed a novel mobile document image retrieval framework. Ruf et al. [85] recognized paintings in art galleries for mobile museum guide. Gui et al. [36] addressed the recognition of large-scale outdoor scenes on smart-phones by fusing outputs of inertial sensors and computer vision techniques. Mouine et al. [74] designed a mobile plant recognition system for plant identification. Auack et al. [84] identified an object from a query image through multiple recognition stages, including local visual features, global geometry, and GPS information. In addition, You et al. [105] proposed a mobile queue-card management system, including store filtering, store recognition and information overlay to enable image-based queue-card retrieving, and service-information querying actions.

5 Databases and performance evaluation

5.1 Databases

In this subsection, we review some representative data sets suitable for mobile visual recognition in the following:

5.1.1 On-premise signs image data set

Tsai et al. [98] released a on-premise signs data set with 62 different businesses categories (OPS-62). This data set totally contains 4649 images from Google's street view, which are common visual objects in our living life. Each image has the pixel-level labeling of the OPS ground truth. This data set is a new benchmark and is suitable for mobile object recognition. In addition, it has been used in on-premise signs recognition [98] and related query management service [105]. The data set is available online[4]

5.1.2 Stanford mobile visual search data set

Chandrasekhar et al. [9] released the Stanford mobile visual search data set for mobile product recognition. The database contains 1 million images, covering different categories: CDs, DVDs, books, software products, landmarks, business cards, text documents, museum paintings, and video clips. They provided a total 3300 query images for 1200 distinct classes across eight image categories. Typically, a small number of query images suffice to measure the performance of a retrieval system as the rest of the database can be padded with "distractor" images. The Stanford mobile visual search data set has been used in many mobile visual search tasks. However, each image from the data does not contain any contextual information. Therefore, this data set is not suitable for the tasks of context-based mobile visual recognition. The data set is available online.[5]

5.1.3 San Francisco data set

Chen et al. [11] released the San Francisco data set for mobile landmark recognition in 2011. The database contains 1.7 million perspective images, where each image is associated with the GPS tag. Data are collected using a mobile mapping vehicle composed of 360° LIDAR sensor, panoramic camera, high-definition cameras, global positioning system (GPS), inertial measurement unit (IMU), and distance measurement instrument (DMI) to obtain panoramas from the San Francisco city. Because the spherical projection alters the locations and the descriptors of local image features, matching query images directly to these panoramas yields poor results. Therefore, these panoramas are further converted into perspective central images (PCIs) and perspective frontal images (PFIs). Each PCI is associated with the field of view, the center of projection, the camera orientation, the visibility mask, and the building label. For each PFI, the warping plane parameters are given. In addition, they released 803 cell phone query images, captured using a variety of mobile phones as part of the data set. Each query image includes GPS information. The San Francisco data set has been used in many mobile visual recognition tasks with GPS contextual information [18, 35, 60, 108]. The entire data set is available online.[6]

5.1.4 Other data set

In addition to these two open data sets, there are also other unpublished data sets suitable for mobile visual recognition. For example, Cheng et al. [24] built a large-scale test collection that consists of (1) 355,141 images about 128 landmarks in five cities over three continents from Flickr and (2) different kinds of textual features for each image, including surrounding text (e.g., tags), contextual data (e.g., geolocation and upload time), and meta-data (e.g., uploader and EXIF). Rong et al. [51] collected over 10 million geo-tagged photos from photo-sharing Websites of Flickr and Panoramio, which covers typical areas (e.g., Beijing, New York City, Singapore, and Florence) for mobile landmark search. Yap et al. [104] created a landmark database consisting of 4000 images of landmarks-50 categories and 80 images per category. They captured the images using different camera phones with different built-in camera settings (e.g., contrast and resolution) from different viewpoints under different weather and illumination conditions. Each image is associated with the GPS and direction information.

5.2 Performance evaluation

For the visual recognition tasks, it is natural to utilize the existing performance metrics of information retrieval (IR) and computer vision to evaluate the performance. In addition, there are also other performance metrics specific for mobile visual recognition, such as the energy consumption and the system latency [33]. In this subsection, we will review these two categories of performance metrics.

5.2.1 Performance metrics for recognition

– *Precision–Recall (P–R curve)* Precision and recall are the traditional metrics in the field of information

[4] http://mclab.citi.sinica.edu.tw/dataset/ops62/ops62.html.
[5] https://purl.stanford.edu/rb470rw0983.
[6] https://purl.stanford.edu/vn158kj2087.

retrieval [6], especially used in retrieval-based mobile visual methods. Precision and recall are actually two metrics, but they are often used together. Precision is the fraction of the retrieved documents in the returned subset, while recall is defined as the fraction of retrieved relevant documents in the whole data set. The simple combined use is the precision–recall curve. Another way of combining these two numbers is via the harmonic mean.

- *Mean Average Precision (MAP)* Another widely adopted performance metric is the average precision over a set of retrieved visual documents. Let $y_i \in \{0, 1\}$ denote if the i-th document d_i in the ranked list r is relevant ($y_i = 1$) or not ($y_i = 0$). The average precision (AP) is defined by $AP = \frac{1}{N} \sum_{i=1}^{N} \frac{y_i}{i} \sum_{j=1}^{i} y_j$, where N is the number of retrieved documents, and $\sum_{j=1}^{i} \frac{y_j}{i}$ is the precision at given rank i. Then, MAP is computed by averaging the AP across all given queries [39, 72].
- *Normalized Discounted Cumulative Gain (NDCG)* NDCG is a commonly adopted metric for evaluating a search engine's performance [46]. NDCG measures the usefulness, or gain, of a ranked list of documents based on their positions in the ranked list.
- *Classification accuracy* This metric is particularly used in classification-based methods. classification accuracy is widely used in classification tasks. Accuracy simply measures how often the classifier makes the correct prediction. It's the ratio between the number of correct predictions and the total number of predictions (the number of test data points). A variation of accuracy is the average per-class accuracy-the average of the accuracy for each class. Accuracy is an example of a micro-average, and average per-class accuracy is an example of a macro-average.

5.2.2 *Performance metrics for mobile platforms*

- *EER* Equal error rate (EER) [8] is the rate, at which both acceptance and rejection errors are equal. In the mobile devices, communication and power costs are significant for transmitting information from the client to the server. Feature compression is, hence, vital for reduction in storage, latency, and transmission. EER is used to compare the quality of different types of compressed features. For the same bit rate, the less the EER, the better the features.
- *System latency* The system latency can be broken down into three components: processing time on client, transmission time, and processing time on server [33]. The transmission time is the time when the data is sent from the client to the server. In [33], the experiment shows that the data transmission time is insignificant for a WLAN network because of the high bandwidth available. However, the transmission time turns out to be a bottleneck for 3G networks. A good mobile visual recognition system requires a low system latency, so that it can satisfy user's interactive experience.
- *Energy consumption* Conserving the energy is critical in mobile visual recognition, because of the limited capacity of the battery. The energy consumption of different mobile devices can be measured in their own platforms, respectively. For example, the average energy consumption associated with a single query on the Nokia 5800 phone is measured using the Nokia energy profiler [33], while the power consumption on the iphone is measured by Apple energy measurement application [73].

In addition to the objective evaluations, there are some subjective evaluations, namely the user experience. For example, Min et al. [73] evaluated their recognition results using the effectiveness and attractiveness. Effectiveness means whether the returned results are correct, while attractiveness means whether the returned results are vivid and attractive.

6 Application scenarios

There is an increasing amount of applications related to CAMVR, and we give a brief introduction in the following.

6.1 Mobile search

Mobile search has made a great contribution to the market. According to a leading market research firm eMarketer,[7] 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",[8] "Kooaba", and Nokia "Point and Find".[9] Google Goggles is a mobile application that lets users search the Web using pictures (e.g., books, artworks, and wine) taken from their mobile phones. Amazon Flow lets 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 display-related information, further links and available files, applied to wine lists,

[7] https://en.wikipedia.org/wiki/Mobile_search

[8] http://flow.a9.com.

[9] http://pointandfind.nokia.com.

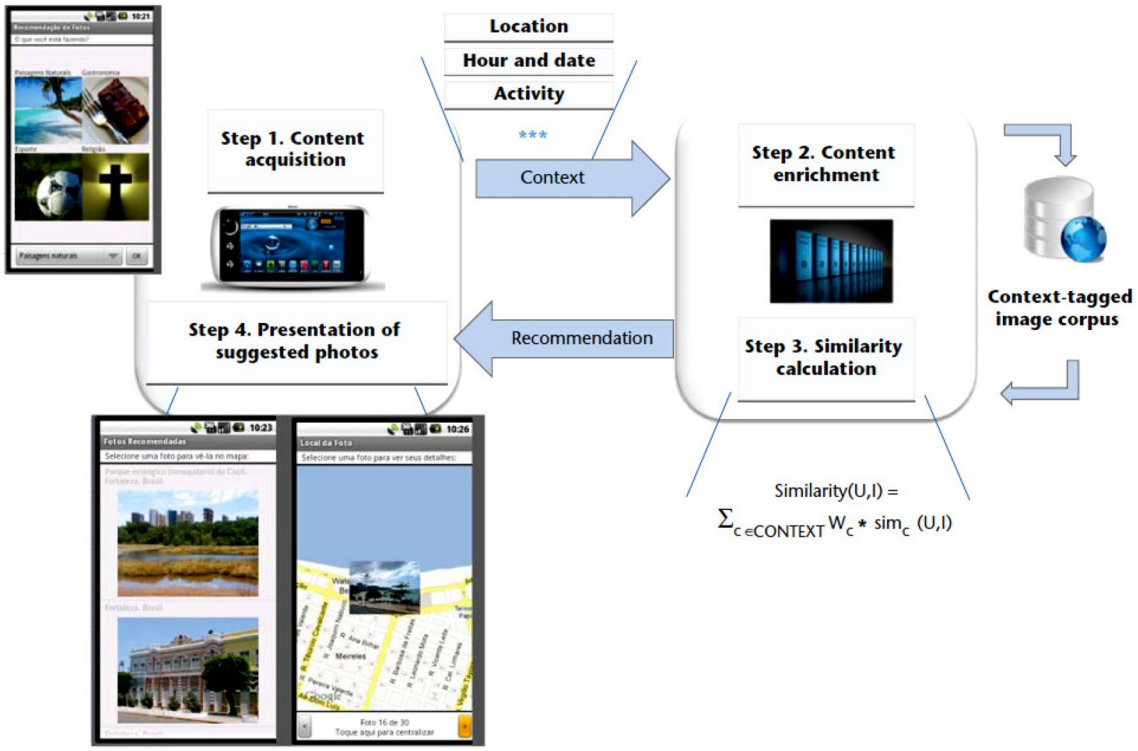

Fig. 8 overview of mobile photo and video recommendation system [99]

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 [33, 40, 96] in academy, such as mobile food recognition [53, 55] and mobile clothes retrieval [28, 29, 68]. In addition, some work [15, 17–22] incorporated the location and direction information to perform mobile landmark search.

6.2 Mobile recommendation

The recognized results can be used for mobile recommendation. For example, mobile landmark recognition [11, 19, 61] can be further used for travel recommendation. Maruyama et al. [71] proposed a mobile cooking recipe recommendation system by recognizing food ingredients, such as vegetables and meats. Zhang et al. [111] allowed a mobile user to take a photo and naturally indicate an object-of-interest within the photo via circle-based 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. [99] proposed a mobile photo and video recommendation system (Fig. 8), which leverages the user's context to enrich and annotate context data, perform a similarity analysis, and provide photo recommendations.

6.3 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[10] offers a shopping service that helps users to 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 [68] proposed a mobile clothes retrieval model: a user takes a photo of any person, then similar clothing from online shops is retrieved using the proposed cross-scenario image retrieval solution to facilitate online clothing shopping. Di et al. [29] proposed an

[10] http://www.omoby.com.

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. [28] 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 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. [106] focused on improving visual search based on mobile shopping experience using machine and crowd intelligence, where the user interaction can be considered as the contextual information.

6.4 Mobile navigation

Mobile visual location search [89] and mobile landmark recognition [15, 17–22] can be used for mobile navigation. Je et al. [47] introduced the street searching service for mobile navigation. The buildings around crossroads are appropriate for 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 search 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 a traditional map and multi-perspective panoramic street views. It can help us search and find out somewhere more intuitively. In addition, Liu et al. [66, 67] presented 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.

6.5 Mobile augmented reality

Mobile augmented reality (MAR) [12] is a wide class of applications where mobile devices augment users' perception of the world. MAR processes a stream of viewfinder frames captured by a mobile device's camera to recognize [33], track, and augment objects that appear in these frames. Chen et al. [14] streamed live videos on the mobile phone to the remote server, on which a SURF-based recognition engine was used to obtain features. In their latest work [12], they developed 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.

6.6 Other potential mobile applications

Mobile visual recognition can also be used for product placement.[11] 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.

7 Conclusions and future research directions

In this survey, we have reviewed the recent work on context-aware mobile visual recognition (CAMVR). We first introduced the available mobile contexts which are commonly used, and showed that the location context is popular for various recognition tasks, and other types of contexts are often used as complementary. Then, we described different recognition methods, and showed that most works 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 contextual information; (2) how to design compact and discriminative descriptors; (3) how to effectively integrate content and contextual information; and (4) how to consider user's intention.

First, compared with general mobile visual recognition, one goal of CAMVR is to utilize rich contextual information to speed up the recognition time and improve the recognition performance. However, the contextual information of most existing works is limited to GPS information or two kinds of contextual information. The constraint of more contextual information can further speeds up the recognition time, and thus, the real-time requirement of mobile visual recognition is more easily satisfied. Therefore, effectively combining more contextual information [44, 75] is desirable.

Second, limited storage capacity and network bandwidth are two limitations of mobile visual recognition. This limits the use of very high-dimensional feature descriptors.

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

Therefore, smaller descriptors with comparable discriminative performance are needed. Although some work [8, 40] have designed compressed descriptors, which achieve almost identical performance as common SIFT descriptors, 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.

Third, most works [11, 19–22] mainly use the contextual 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 contextual information probably further improves the recognition performance. Some works [18, 51] utilized the contextual information (e.g., GPS and direction information) in the offline discriminative feature learning and improved the recognition performance. Therefore, it would be interesting to integrate more contextual information to the content information for more effective feature learning in the offline phase.

Finally, since mobile visual recognition should address user's needs, the ideal mobile visual recognition should take the user intent into account. Few work [112] considered the user intent in mobile visual recognition. Therefore, how to incorporate the user intent into mobile visual recognition is probably an interesting research topic.

Acknowledgments This work was supported in part by the National Basic Research 973 Program of China under Grant No. 2012CB316400, the National Natural Science Foundation of China under Grant Nos. 61532018, 61322212, 61303160, 61572488 and 61550110505, China Postdoctoral Science Foundation under Grant No. 2016M590135, the National High Technology Research and Development 863 Program of China under Grant No. 2014AA015202. This work is also funded by Lenovo Outstanding Young Scientists Program (LOYS).

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. Arandjelović, R., Zisserman, A.: Name that sculpture. In: Proceedings of the 2nd ACM International Conference on Multimedia Retrieval, p. 3. ACM (2012)
4. Arandjelović, R., Zisserman, A.: Three things everyone should know to improve object retrieval. In: Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, pp. 2911–2918. IEEE (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. Baeza-Yates, R., Ribeiro-Neto, B., et al.: Modern information retrieval, vol. 463 (1999)
7. Bay, H., Tuytelaars, T., Van Gool, L.: Surf: Speeded up robust features. In: Computer vision—ECCV 2006, pp. 404–417. Springer, Berlin (2006)
8. 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)
9. Chandrasekhar, V.R., Chen, D.M., Tsai, S.S., Cheung, N.M., Chen, H., Takacs, G., Reznik, Y., Vedantham, R., Grzeszczuk, R., Bach, J., et al.: The stanford mobile visual search data set. In: Proceedings of the Second Annual ACM Conference on Multimedia Systems, pp. 117–122. ACM (2011)
10. Chatzilari, E., Liaros, G., Nikolopoulos, S., Kompatsiaris, Y.: A comparative study on mobile visual recognition. In: Machine Learning and Data Mining in Pattern Recognition, pp. 442–457. Springer, Berlin (2013)
11. 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)
12. 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)
13. 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)
14. 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)
15. 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)
16. 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)
17. Chen, T., Lu, S., Fan, J.: Context-aware vocabulary tree for mobile landmark recognition. J. Vis. Commun. Image Represent. **30**, 289–298 (2015)
18. Chen, T., Yap, K.H.: Context-aware discriminative vocabulary learning for mobile landmark recognition. Circuits Syst. Video Technol. IEEE Trans. **23**(9), 1611–1621 (2013)
19. Chen, T., Yap, K.H.: Discriminative bow framework for mobile landmark recognition. Cybern. IEEE Trans. **44**(5), 695–706 (2014)
20. 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)
21. Chen, T., Yap, K.H., Chau, L.P.: Integrated content and context analysis for mobile landmark recognition. Circuits Syst. Video Technol. IEEE Trans. **21**(10), 1476–1486 (2011)
22. Chen, T., Yap, K.H., Zhang, D.: Discriminative soft bag-of-visual phrase for mobile landmark recognition. Multimed. IEEE Trans. **16**(3), 612–622 (2014)

23. 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)
24. Cheng, Z., Ren, J., Shen, J., Miao, H.: Building a large scale test collection for effective benchmarking of mobile landmark search. In: Advances in Multimedia Modeling, pp. 36–46. Springer, Berlin (2013)
25. Chi, H.Y., Chen, C.C., Cheng, W.H., Chen, M.S.: Ubishop: commercial item recommendation using visual part-based object representation. Multimed. Tools Appl. pp. 1–23 (2015)
26. Chi, H.Y., Cheng, W.H., Chen, M.S., Tsui, A.W.: Mosro: Enabling mobile sensing for realscene objects with grid based structured output learning. In: International Conference on Multimedia Modeling, pp. 207–218. Springer (2014)
27. Chum, O., Mikulik, A., Perdoch, M., Matas, J.: Total recall II: query expansion revisited. In: Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on, pp. 889–896. IEEE (2011)
28. 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)
29. 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)
30. 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 Process. Image Commun. **28**(4), 368–385 (2013)
31. Duan, L.Y., Ji, R., Chen, Z., Huang, T., Gao, W.: Towards mobile document image retrieval for digital library. Multimed. IEEE Trans. **16**(2), 346–359 (2014)
32. 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)
33. 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 Process. Mag. IEEE **28**(4), 61–76 (2011)
34. Guan, T., He, Y., Duan, L., Yang, J., Gao, J., Yu, J.: Efficient bof generation and compression for on-device mobile visual location recognition. MultiMed. IEEE **21**(2), 32–41 (2014)
35. Guan, T., He, Y., Gao, J., Yang, J., Yu, J.: On-device mobile visual location recognition by integrating vision and inertial sensors. Multimed. IEEE Trans. **15**(7), 1688–1699 (2013)
36. Gui, Z., Wang, Y., Liu, Y., Chen, J.: Mobile visual recognition on smartphones. J. Sens. **2013**, 1–9 (2013)
37. Guillaumin, M., Mensink, T., Verbeek, J., Schmid, C.: Tagprop: Discriminative metric learning in nearest neighbor models for image auto-annotation. In: Computer Vision, 2009 IEEE 12th International Conference on, pp. 309–316. IEEE (2009)
38. Hao, J., Wang, G., Seo, B., Zimmermann, R.: Point of interest detection and visual distance estimation for sensor-rich video. Multimed. IEEE Trans. **16**(7), 1929–1941 (2014)
39. Hauptmann, A.G., Christel, M.G.: Successful approaches in the trec video retrieval evaluations. In: Proceedings of the 12th Annual ACM International Conference on Multimedia, pp. 668–675. ACM (2004)
40. 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)
41. 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)
42. Herranz, L., Xu, R., Jiang, S.: A probabilistic model for food image recognition in restaurants. In: Proceedings of the IEEE ICME (2015)
43. Houle, M.E., Oria, V., Satoh, S., Sun, J.: Annotation propagation in image databases using similarity graphs. ACM Trans. Multimed. Comput. Commun. Appl. (TOMM) **10**(1), 7 (2013)
44. 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)
45. 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)
46. Järvelin, K., Kekäläinen, J.: Ir evaluation methods for retrieving highly relevant documents. In: Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 41–48. ACM (2000)
47. 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)
48. 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)
49. 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)
50. 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)
51. Ji, R., Duan, L.Y., Chen, J., Yao, H., Yuan, J., Rui, Y., Gao, W.: Location discriminative vocabulary coding for mobile landmark search. Int. J. Comput. Vis. **96**(3), 290–314 (2012)
52. 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)
53. 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)
54. 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)
55. Kawano, Y., Yanai, K.: Foodcam: a real-time mobile food recognition system employing fisher vector. In: MultiMedia Modeling, pp. 369–373. Springer, Berlin (2014)
56. 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)
57. 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)
58. 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)
59. 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)
60. Li, Z., Yap, K.H.: Content and context boosting for mobile landmark recognition. Signal Process. Lett. IEEE **19**(8), 459–462 (2012)
61. Li, Z., Yap, K.H.: Context-aware discriminative vocabulary tree learning for mobile landmark recognition. Digital Signal Process. **24**, 124–134 (2014)
62. Li, Z., Yap, K.H., Tan, K.W.: Context-aware mobile image annotation for media search and sharing. Signal Process. Image Commun. **28**(6), 624–641 (2013)
63. 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)
64. Lin, J., Wu, V.: Tagging content with metadata pre-filtered by context (2013). https://www.google.com/patents/US8370358. US Patent 8,370,358
65. Liu, H., Li, H., Mei, T., Luo, J.: Accurate sensing of scene geo-context via mobile visual localization. Multimed. Syst. **21**(3), 255–265 (2015)
66. Liu, H., Mei, T., Li, H., Luo, J., Li, S.: Robust and accurate mobile visual localization and its applications. ACM Trans. Multimed. Comput. Commun. Appl. (TOMM) **9**(1s), 51 (2013)
67. 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)
68. 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)
69. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vis. **60**(2), 91–110 (2004)
70. Mai, W., Dodds, G., Tweed, C. (eds.): A pda-based system for recognizing buildings from user-supplied images. In: Mobile and Ubiquitous Information Access, pp. 143–157. Springer, Berlin (2004)
71. 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)
72. Mei, T., Rui, Y., Li, S., Tian, Q.: Multimedia search reranking: a literature survey. ACM Comput. Surv. (CSUR) **46**(3), 38 (2014)
73. Min, W., Xu, C., Xu, M., Xiao, X., Bao, B.K.: Mobile landmark search with 3d models. Multimed. IEEE Trans. **16**(3), 623–636 (2014)
74. 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)
75. Naaman, M., Nair, R.: Zonetag's collaborative tag suggestions: What is this person doing in my phone? MultiMed. IEEE **15**(3), 34–40 (2008)
76. 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, Berlin (2003)
77. Nister, D., Stewenius, H.: Scalable recognition with a vocabulary tree. In: Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on, vol. 2, pp. 2161–2168. IEEE (2006)
78. O'Hare N., Gurrin C., Jones G.J., Smeaton A.F. Combination of content analysis and context features for digital photograph retrieval. In: Proceedings of 2nd IEE European Workshop on the Integration of Knowledge, Semantic and Digital Media Technologies, pp. 323–328, IEEE Computer Society, London, UK, Washington, DC, USA, November 29–December 1, 2005
79. Oliveira, L., Costa, V., Neves, G., Oliveira, T., Jorge, E., Lizarraga, M.: A mobile, lightweight, poll-based food identification system. Pattern Recognit **47**(5), 1941–1952 (2014)
80. 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)
81. 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)
82. 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)
83. Qin, C., Bao, X., Choudhury, R.R., Nelakuditi, S.: Tagsense: leveraging smartphones for automatic image tagging. Mob. Comput. IEEE Trans. **13**(1), 61–74 (2014)
84. Quack, T., Bay, H., Van Gool, L.: Object recognition for the internet of things. In: Floerkemeier, C., Langheinrich, M., Fleisch, E., Mattern, F., Sarma, S.E. (eds.) The Internet of Things, pp. 230–246. Springer, Berlin (2008)
85. Ruf, B., Kokiopoulou, E., Detyniecki, M.: Mobile museum guide based on fast sift recognition. In: Detyniecki, M., Leiner, U., Nürnberger, A. (eds.) Adaptive Multimedia Retrieval. Identifying, Summarizing, and Recommending Image and Music, pp. 170–183. Springer, Berlin (2010)
86. 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 and Services, pp. 513–518. ACM (2014)
87. Sang, J., Mei, T., Xu, Y.Q., Zhao, C., Xu, C., Li, S.: Interaction design for mobile visual search. Multimed. IEEE Trans. **15**(7), 1665–1676 (2013)
88. 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)
89. Schroth, G., Huitl, R., Chen, D., Abu-Alqumsan, M., Al-Nuaimi, A., Steinbach, E.: Mobile visual location recognition. Signal Process. Mag. IEEE **28**(4), 77–89 (2011)
90. 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, Berlin (2004)
91. 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, Berlin (2012)
92. 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)

93. 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)
94. 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)
95. 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)
96. 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)
97. 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)
98. Tsai, T.H., Cheng, W.H., You, C.W., Hu, M.C., Tsui, A.W., Chi, H.Y.: Learning and recognition of on-premise signs from weakly labeled street view images. Image Process. IEEE Trans. **23**(3), 1047–1059 (2014)
99. 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. MultiMed. IEEE **21**(1), 24–34 (2014)
100. 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)
101. Xie, X., Lu, L., Jia, M., Li, H., Seide, F., Ma, W.Y.: Mobile search with multimodal queries. Proc. IEEE **96**(4), 589–601 (2008)
102. Xu, R., Herranz, L., Jiang, S., Wang, S., Song, X., Jain, R.: Geolocalized modeling for dish recognition. Multimed. IEEE Trans. **17**(8), 1187–1199 (2015)
103. 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)
104. Yap, K.H., Chen, T., Li, Z., Wu, K.: A comparative study of mobile-based landmark recognition techniques. Intell. Syst. IEEE **25**(1), 48–57 (2010)
105. You, C.W., Cheng, W.H., Tsui, A.W., Tsai, T.H., Campbell, A.: Mobilequeue: an image-based queue card management system through augmented reality phones. In: Proceedings of the 2012 ACM Conference on Ubiquitous Computing, pp. 651–652. ACM (2012)
106. 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)
107. 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)
108. 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)
109. Zamir, A.R., Dehghan, A., Shah, M.: Visual business recognition: a multimodal approach. In: ACM Multimedia, pp. 665–668. Citeseer (2013)
110. 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). doi:10.1016/j.neucom.2014.10.105
111. 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)
112. 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)