IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received May 6, 2021, accepted June 15, 2021, date of publication June 29, 2021, date of current version July 28, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3093462

Rich Common Crucial Feature for Crowdsourcing-Based Mobile Visual Location Recognition

HAO WANG[1], YUGUI WANG[2], RUI CUI[3], YIBO HAN[4], CHAOHUA YAN[1], AND MENGHAN NIU[1]

[1]School of Computer and Information Engineering, Nanyang Institute of Technology, Nanyang, Henan 473004, China
[2]School of Information Engineering, Nanjing Polytechnic Institute, Nanjing 210018, China
[3]School of Computer and Information Technology, Nanyang Normal University, Nanyang, Henan 473061, China
[4]Institute of Bigdata, Nanyang Institute of Technology, Nanyang, Henan 473004, China

Corresponding author: Rui Cui (cuiruinynu@163.com)

This work was supported in part by the Scientific and Technological Project in Henan Province under Grant 212102210384, in part by the Key Scientific Research Project of Colleges and Universities in Henan Province under Grant 21A520031, in part by the Nanyang Institute of Technology through the Interdisciplinary Sciences Project under Grant 520019, in part by the Scientific and Technological Project in Nanyang City under Grant KJGG007, and in part by the Education and Teaching Reform Project under Grant NIT2020JY-077.

ABSTRACT Crowdsourcing provides an effective way to construct a location recognition image database. Comparing with traditional location image database construction, crowdsourced image database has massive advantages, e.g., much richer information for location, with various angles, timestamps, distances and weather information, providing useful potential for high recognition precision. However, when capturing the crowdsourced images, it is inevitable to have various disturbances on these location images, for example, moving vehicles and pedestrians, hindering the realization of potential. To address this challenge, we first propose a Rich Common Crucial Feature (RCCF) detection framework to exclude unimportant visual SURF features from crucial features. To achieve a good balance between the efficiency and accuracy, we further propose an RCCF based Visual Hash Bits (VHB) scheme to encode RCCF features into hash bits to vote for most matching images. Furthermore, deep feature extraction is also utilized with visual search architecture MobileNet. Extensive experiments are conducted on a crowdsourced dataset with 9,064 location images, demonstrating that our scheme outperforms other state-of-the-art schemes.

INDEX TERMS Crowdsourcing, rich information, deep hash feature, common crucial feature.

I. INTRODUCTION

With the ubiquitous smartphones, it is convenient to perform various location-based services, for example, location recognition systems. In these systems, GPS and Wi-Fi based localization methods have been widely used in outdoor environments [1], [2]. Although these systems are capable of capturing the preliminary physical coordinates (GPS information), it is poorly-recognized with respect to the logical meanings of scenes, e.g., buildings, landmarks, shops that the users are interested in. Moreover, the localization accuracy is still not satisfactory in many realistic scenarios. For example, errors of localization of GPS often [1] range from 5 to 300 meters at some places with poor visibility. Comparing with these location recognition systems, the Mobile Visual Location Recognition (MVLR) [3] combines captured image and sensory data obtained from smartphone as a location query, which provides logical location information as an important supplement.

Constructing location image database is one basic prerequisite for MVLR. One way is to collect location images from online photo sharing sites. For instance, the Oxford Buildings Dataset [4] consists of 5,062 images collected from Flickr, i.e., the above images in Fig. 1. However, these images tend to be disorganized, poorly labeled, and unevenly distributed in the real world [5]. Another way is to utilize moving vehicles with a digital camera system to capture panorama images. However, this way is labor-intensive, time-consuming, and

The associate editor coordinating the review of this manuscript and approving it for publication was Grigore Stamatescu.

even useless wherever vehicles cannot reach, e.g., restricted areas in some business streets [5]. In order to address this issue, we propose a crowdsourcing based framework in the previous works [5]–[9] to construct a location image database, e.g., the images in Fig. 1.

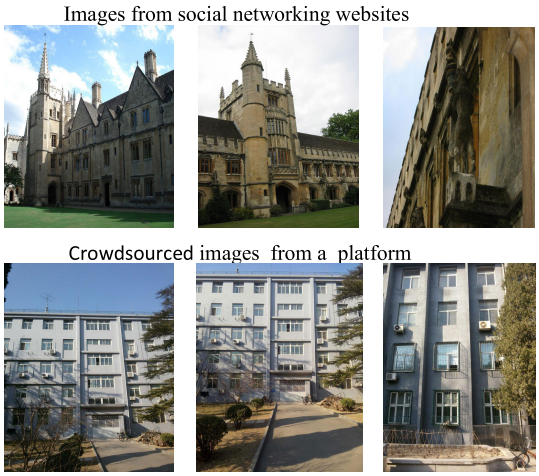

FIGURE 1. Comparison of images from social networking websites and images from a crowdsourcing platform.

Comparing with Oxford Buildings Dataset, we can find that the view direction and distance of crowdsourced images are more suitable for matching actual location query image as shown in Fig. 1. Specifically, some smartphone users are recruited (possibly with some incentives [5]–[9]) to collect location images from different distances and angles for constructing a fundamental database. Meanwhile, with the increase of user query images, the database will be updated continuously. More importantly, the crowdsourced database has rich information captured with various angles, distances, times and weathers, and is always well organized and well-labeled, so it has greater potential for high recognition accuracy.

Nevertheless, as shown in Fig. 2, we still have a big problem with respect to the quality of crowdsourced images [9], because it is inevitable to have various disturbances on these location images. These disturbances may be moving pedestrians, vehicles and their shadows; or may be caused by changing seasons (e.g., flourishing leaves in summer will fall in autumn) and weathers (e.g., with rains or snows); or may be different trivial objects near the object of interest captured from different distances and angles. The challenge lies in how to take full advantage of the rich information of crowdsourced images and decrease negative effects of the aforementioned disturbances. As shown in Fig. 2, disturbance parts are presented by triangles while the crucial part features are presented by red points and linked by lines. From these images, we can find that crucial part features are more robust, compared with other features.

To address the problem, we propose a **R**ich information based **C**ommon **C**rucial **F**eature (RCCF) detection framework to assign each **S**peeded **U**p **R**obust **F**eatures (SURF) [16] feature a weight in the image database. RCCF detection is achieved by determining the best matching pairs between the images with the same object. Most importantly, the rich information refers to the sensory data, such as physical coordinate, view direction, tilt angle, etc. Inspired by ISP [10], we first exploit the Visual Hash Bits (VHB) scheme [11] to compress the features into binary codes, because of the good balance between the efficiency and search accuracy of VHB in mobile visual search. Second, to further improve the recognition accuracy of VHB, we propose a RCCF based VHB scheme, which assigns every RCCF feature a weight value learned from the crowdsourced image database. Then, we also transfer the rich information to both Bag of Words (BOW) and MobileNet method, using deep learning hash to testify the RCCF. Finally, we focus on decreasing the impacts of disturbances in a crowdsourced image database. It is noted that although not all the disturbance features are excluded, our scheme indeed distinguishes the common and crucial features from others in the crowdsourced database to decrease the undesirable impacts of the disturbance. It is concerned that the crucial and robust features in the crowdsourced image database is used to decrease the undesirable impacts of the non-common crucial features, as illustrated in Fig. 2.

FIGURE 2. Examples of disturbances in the crowdsourced images.

In this paper, we propose a RCCF detection algorithm, which is an efficient method to exclude disturbances of crowdsourced images by the rich information. Comparing with the traditional location recognition algorithms, RCCF realizes the following benefits: First, the algorithm of the background is ubiquitous for the crowdsourced database, which has been widely used in outdoor environments. Second, the method utilizes the rich information to achieve location recognition accuracy improvement in the crowdsourced databases through extensive experiments. Specifically, we conduct the experiments on a crowdsourced location image dataset(BUPT Dataset) [6], [8], [9], which is composed of 9,064 images crowdsourced from 172 objects.

In summary, the main contributions of this paper are summarized as follows:

- We utilize the rich sensory information for crowdsourcing database disturbance exclusion problem.

- To achieve disturbance exclusion goal, we propose RCCF detection algorithm, in which the best matching pairs are designed and the each weight of the features is assigned.
- Experimental experiments are conducted to show that our scheme outperforms other state-of-the-art schemes, in terms of the widely used performance evaluation metrics.

The rest of the paper is organized as follows. In Section II, we summarize the related work about crowdsourced image selection and common crucial feature detection technology. In Section III, we describe the motivation and framework of MVLR. In Section V, we present the spatial division with respect to the sensory data associated with the images. In Section VI, we present crucial part feature detection approach for evaluating the image quality. In Section VII, extensive experiments are conducted to demonstrate the benefits of the framework. At last, the paper is concluded in Section VIII.

II. RELATED WORK

In the section, we discuss the related works about crowdsourcing based image collection and rich infirmation detection on common feature approaches.

A. CROWDSOURCING-BASED IMAGE COLLECTION

In recent years, abundant researchers focus on crowdsourcing technology to perform specific tasks consciously or unconsciously. Actually, researchers deploy different forms of individual Human Intelligence Task (HIT) on crowdsourcing platform to collect various images.

We illustrate different forms of HIT with respect to the crowdsourcing technology. For instance, Google images [13] introduce an interactive game, in which the players unconsciously help determine the contents of images. With the form of HIT, a crowd of people is recruited to determine the contents of images through providing meaningful labels. For another type of HIT, Rudinac *et al.* propose MTurk [12] system, which has been used as powerful tools for implementing relatively efficient image collection task. MTurk system recruits participants to select the representative images capable of indicating the meaningful contents. In this form of HIT, MTurk succeeds in selecting ten images from a given 100-image set, which is able to give the participants an overall impression. For the form of HIT in our previous works [5]–[7], CrowdOLR system asks the participants to capture the objects (buildings, statues, Landmarks etc.) with various of view direction and distance. Therefore, the crowdsourced BUPT dataset [5] has more rich information, which is capable of depicting precise geo-annotations. However, it is inevitable that these crowdsourced image database is collected with massive noisy.

B. COMMON FEATURE DETECTION WITH RICH INFORMATION

Extensive research has been conducted on common or salient SURF detection, which refers to finding salient features in the images corresponding to the same object. For example, Yang *et al.* [10] propose Identical Salient Point (ISP) to extract salient visual words from multiple images captured by smartphones, using which some disturbances are excluded for improving searching precision. Wang *et al.* [9] propose Salient Part Feature Detection(SPFD) algorithm to distinguish the salient part features from others to evaluate the image quality of crowdsourced database within the spacial clusters. Meanwhile, Chen *et al.* [14] design a generic task-driven data collection and selection framework (e.g., CrowdPic) to meet diverse mobile crowd photographing application requirements with respect to multiple sensory data. Wang *et al.* [15] propose an image selection framework SmartPhoto, which evaluates the contribution of crowdsourced photos according to geographical and geometrical data. Whereas, our goal is to propose the unimportant visual feature exclusion method from crucial features, decreasing the redundant disturbances to reduce the burden of storage by using the rich information of the crowdsourced database.

III. MOTIVATION AND FRAMEWORK

In this section, we first illustrate our motivation, and then describe the framework as shown in Fig. 3.

A. MOTIVATION

As mentioned before, it is inevitable that the crowdsourced images have various disturbances, e.g., moving pedestrians, vehicles and their shadows. However, it has been testified that these disturbances can be excluded to improve searching precision, using salient visual words [10] from multiple images captured by smartphone. Inspired by [10], it is observed that the features corresponding to the disturbances are not always shared in the images with the same object. Whereas, the common and crucial parts are shared by two images with the same object. In other words, non-crucial parts are more possibly to be missed. Therefore, we propose RCCF method to exclude the disturbances, in which the non-crucial part determination method will be illustrated later.

Inspired by ISP [10], we focus on contextual saliency determination from the multiple relevant images corresponding to the same object, which is more robust, stable and significant. Following ISP [10], we determine RCCF by matching the best matching pair [9] between every several relevant images to capture common contents. RCCF features are determined by the best match pairs, which belong to the images with the same object. Specifically, suppose *imageA* have feature set $F : [F1, F2, \ldots]$, and *imageB* have feature set $F' : [F1', F2', \ldots]$. Each feature is represented by 80 binary bit codes as $[c1, c2, \ldots c_{80}]$. We determine the best matching pairs as follows: for each feature f in the F, we match it with every feature in the set F' and obtain the best matching feature f' with the minimum distance; Meanwhile, we match f' with each feature in F. If the best matching feature is also f, we regard (f, f') as a best match pair and both of them are RCCF features.

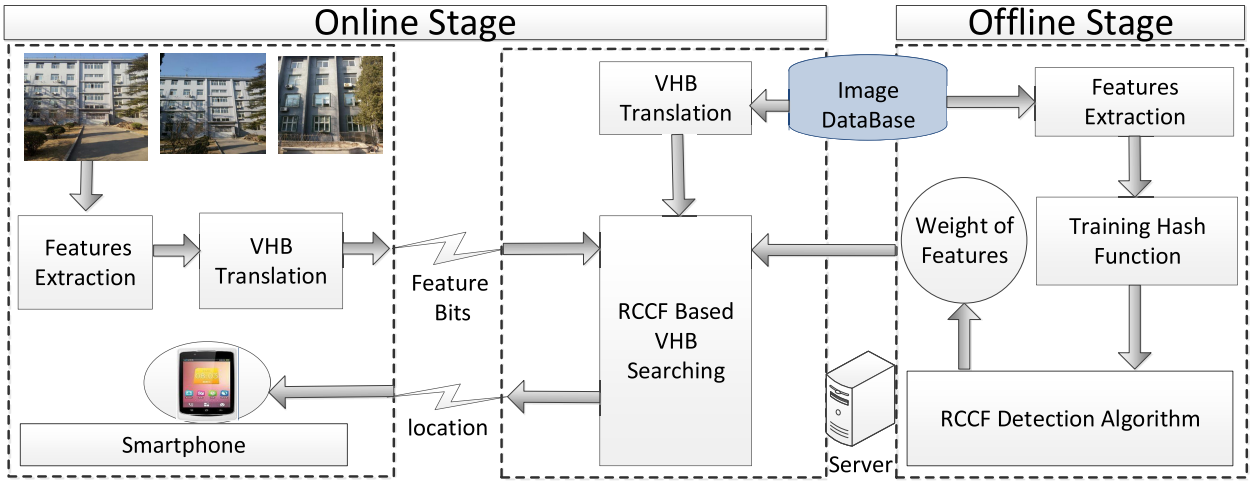

FIGURE 3. Framework of rich common crucial feature is composed of online stage and offline stage.

B. FRAMEWORK

Fig. 3 shows the framework of RCCF detection, which is composed of two parts: offline stage and online stage.

In the offline stage, the main task is training the weight of extracted features. First, we extract SURF features from the images of BPUT dataset. Next, the features are utilized to train the hash function, which is used to convert these SURF features into binary hash codes. Finally, we propose RCCF detection algorithm, which is used to assign weights of every feature in the training dataset. Alternately, we also train the image dataset by Bag of Words (BOW) method and determine matching SIFT points pairs according to ISP [10], with respect to the BOW visual words. we also calculate their distances (u, q) by Eq. (1).

$$Dis(u, q) = (u.q)/(|u|.|q|) \tag{1}$$

In Eq. (1), u and q denoate 128D SIFT descriptor vector and $|u|$ is the norm of vector u. ISP points are determined by the matching score, which is denoted as:

$$ISP_l = \{dis_l^1, dis_l^2, \ldots, dis_l^i, \ldots, dis_l^V\} \tag{2}$$

In Eq. (2), V is the dimension of vector u and q.

In the online stage, we also extract SURF features from query image and convert them into binary codes, which are sent to the server by wireless networks. On the server, RCCF based VHB image searching method is conducted, and return the best results to the smartphone.

IV. RICH COMMON CRUCIAL FEATURE DETECTION

In this section, we first describe the common crucial feature detection method with rich information, including training process of hash function. Then, we also consider the BOW based RCCF detection method.

A. COMMON CRUCIAL FEATURE DETECTION

Inspired by successful utilization of VHB in mobile visual search [17], we adopt the Spectral Hash [18] method to encode the SURF feature into a sequence of binary hash codes in our system. The offline training stage consists of the following steps: Firstly, for all images in the database, we extract their SURF features as its good balance between the efficiency and search accuracy in mobile visual search [19]. Next, the Spectral Hash is used to unsupervisely learn the hash function shown in equation 6. After that, with the learned transition parameter matrix W, we can transform the 64-dimension float SURF feature into 80 bits binary hash codes. The binary hash codes can decrease the feature transmission and search time of mobile location system [11].

$$h^v = sign(\cos(W\phi)) \tag{3}$$

In next section, we use RCCF detection algorithm to extract the weights of all the features for the RCCF based VHB searching in the online stage.

Recently Identical Salient Point (ISP) [2], [10] scheme is proposed to extract silent visual words from multiple images on smartphone end. However, ISP aims at extracting identical visual words using Bag of Word (BOW) method to exclude the disturbances, without considering the characteristics of crowdsourced images. Moreover, multiple images have to be required. Inspired by the ISP, among the features labeled with feature ID in the crowdsourced image database, our goal in the subsection is to extinguish the RCCF from the non-RCCF.

For the image training set in the server end, we build a object set $O = \{O_1, O_2, \ldots, O_i, \ldots\}$. Then, for each element in O_i, we build a image list vector $O_i = \{I_1, I_2, \ldots, I_j, \ldots\}$, where I_j means the *jth* image belongs to the O_i in the image database. I_j is denoted by feature set $\{F_1, F_2, \ldots., F_k, \ldots\}$. Among the features in the images with the same object, we determine RCCF features by determining the best match pairs. Specifically, suppose image A have feature set $F : [F1, F2, \ldots]$, and *imageB* have feature set $F' : [F1', F2', \ldots]$. For each feature, BOW visual word or 80 binary bit codes are denoted as $[c1, c2, \ldots c_{80}]$.

The best matching pairs determination is composed of two forms, for BOW visual word or hash bit. In the form of hash bit, it is conducted that for each feature f in the F, we match it with every feature in the set F' and obtain the best matching feature f' with the minimum distance; Then, we match f' with each feature in F. If the best matching feature is also f, (f, f') are regarded as a best match pair and both of them are RCCF features. In other words, the best match pair (f, f') satisfies the equation 4.

$$\begin{cases} argmin_{x \in F'}\{Dis(f, x)\} = f' \\ argmin_{x \in F}\{Dis(x, f')\} = f \end{cases} \tag{4}$$

the hamming distance between binary code features is determined by the equation 5.

$$Dis(f, f') = \sum_{1}^{k=80} c_i^k \bigoplus c_j^k \tag{5}$$

Most importantly, we determine the images which have the best matching pairs, according to the spacial similarity by the rich sensory data. The basic idea is that the closer of spacial distance the images have, the more similar the images are. With respect to the spacial distance, we consider the rich information, which is derived from the crowdsourced database. For the same object, the physical coordinate (Longitude and latitude) is obtained from GPS/GSM/WiFi. The view direction and tilt angle are obtained from Accelerometer and Magnetometer [8]. With respect to the spatial distance, we first consider the spatial cluster number to determine the most similar images. Suppose K is the spatial cluster number, which should satisfy the Eq. (6).

$$K = \left\lceil \frac{|Ang_{max} - Ang_{min}|}{20} \right\rceil \tag{6}$$

In Eq. 6, Ang_{max} and Ang_{min} denotes the max tilt angle and min tilt angle.

After determine the cluster number K, we should consider that how to adapt the spatial distribution to clustering number. For the different objects, these images may have different spatial distribution, it is reasonable to assume that the larger the shooting angle differences have, the more cluster number should have. Specifically, if the images of an object occupy a large spatial scope, the clustering number should be designed larger; whereas, if the images of an object occupy a small spatial scope, the clustering number should be designed smaller. Inspired by Crowd-pan-360 [1], we set each shooting scope of cluster to be 20^o as angular threshold parameter. Moreover, from the experimental results in CrowdLR [8], it is also inferred that 20^o is also an appropriate threshold.

After determining the clustering number of the scope of capturing the objects, widely-used K-means algorithm is utilized to the group of images into K clusters. Finally, every image is assigned a unique cluster ID, using which the scope of capturing the images are obtained. We also assign every image the distance scale formulated later with respect to the distance between participant and object.

Algorithm 1: RCCF Detection Algorithm

Input: a object set $O = \{O_1, O_2, \ldots, O_i, \ldots\}$
Output: $W : [w1, w2, \ldots, wn]$

```
foreach i ← 1 to |O| do
    foreach j ← 1 to |O_i| − 1 do
        foreach l ← j + 1 to |O_i| do
            if f ∈ image[j] and f' ∈ image[l]
               satisfy eq.(4, 5, 6) then {
                  w_p ← 1;
                  w_q ← 1;
            }
            /* p corresponds to f ID;
            q corresponds to f' ID */ ;
        end
    end
end
```

To assign the weight of every feature, we assign 1 to the weight of the feature if the feature belongs to RCCF in the training database; vice versa, if the feature belongs to non-RCCF, we assign 0 to the weight of the feature. Then, we obtain the weight vector $W : [w1, w2, \ldots, wn]$, n is the number of the features in the training image database. We summary the RCCF Detection algorithm as Algorithm 1.

B. VISUAL BAG OF WORD RCCF DETECTION

In the form of BOW, we also follow the algorithm as Algorithm 1. BOW based RCCF Detection Algorithm is shown in Algorithm 2, in which BOW algorithm is also a wildly-used image search method and the spatial relationship of visual words is considered. We conduct the spatial RCCF detection method as follows:

- we extract all the SURF features in the image training set to calculate the codebook by Vocabulary Tree, which is essentially hierarchical clustering method. Then, the BOW histogram of each image is obtained by the visual sorted list. Meanwhile, non-visual index list is also obtained by spatial partition, which is constructed by the rich information of the crowdsourced database. The distance of visual words V and U which are presented as vectors is calculated in Eq. 7.

$$DisVU(V, U) = \frac{\sum_{i=1}^{N} V_i \times U_i}{\sqrt{\sum_{i=1}^{N} V_i^2} \times \sqrt{\sum_{i=1}^{N} U_i^2}} \tag{7}$$

- The MVLR APP users send image query data which is composed of sensory and compressed image captured by various sensors of smartphones, e.g., Accelerometer and Magnetometer. Moreover, the MVLR APP users send the compressed data through wireless networks to the MVLR system server.
- On the server, the image query data will be decomposed and compared the similarity of image and sensory data. Image will be translated into BOW histogram by the

codebook to determine the image similarity; meanwhile, the location For instance, the spatial coordinate of the data will be calculated by determining the most similar the image and sensory data. Furthermore, the results will be returned.

$$\begin{cases} argmin_{x\in V}\{DisVU(f,x)\} = f' \\ argmin_{x\in U}\{DisVU(x,f')\} = f \end{cases} \tag{8}$$

the distance between binary code features is determined by the equation 7.

Algorithm 2: BOW Based RCCF Detection Algorithm

Input: a object set $O = \{O_1, O_2, \ldots, O_i, \ldots\}$
Output: $W : [V1, V2, \ldots, Vn]$
// V denotes visual word
foreach $i \leftarrow 1$ *to* $|O|$ **do**
  foreach $j \leftarrow 1$ *to* $|O_i| - 1$ **do**
    foreach $l \leftarrow j + 1$ *to* $|O_i|$ **do**
      if $f \in image[j]$ *and* $f' \in image[l]$
      satisfy eq.(7, 8) **then** {
        $V_p \leftarrow 1$;
        $V_q \leftarrow 1$;
      }
      /* p corresponds to f ID;
      q corresponds to f' ID */ ;
    end
  end
end

C. DEEP HASH METHOD

With the proliferation of deep learning, we also try to use the rich information based RCCF method to search the best matching objects. For simplicity, we utilize the MobileNet system, whose architecture significantly increases the search accuracy with deep feature extraction. Specifically, we adopt the MobileNet with a hash-like layer and train the model on the crowdsourced dataset. We try to automatically train effective binary code feature from the crowdsourced image database using a deep neural network. For one thing, the mobile visual search architecture MobileNet is useful for solving the efficient search on the mobile end. For another, the high computational complexity of the deep neural network is effectively deduced.

MobileNet model is constructed on the depthwise neural convolutions, which utilize a standard convolution and a 1×1 convolution. The standard convolutional layer has input feature map F as $D_F \times D_F \times M$, and the output feature map G as $D_F \times D_F \times N$. D_F denotes the spatial width and height of input feature map with the input depth as M channels. D_G denotes the spatial width and height of input feature map with the output depth as N channels. The standard computational cost of convolution is represented as Eq. 9.

$$D_K \times D_K \times M \times N \times D_F \times D_F \tag{9}$$

To save computation cost, MobileNet use 3×3 depthwise separable convolutions with only a small reduction in accuracy.

For the hash method, the hash codes should have semantic similarity between image label. Specifically, the image with sensory data should be mapped to similar binary codes.

$$D_K \times D_K \times M \times N \times D_F \times D_F \tag{10}$$

Evaluations show that the proposed system can exceed state-of-the-art accuracy performance in terms of the MAP. Moreover, the memory consumption is much less than other deep learning models.

Hash Function is essential to learn from images, which is based on several principles. First, the hash codes should be subordinated to semantic similarity among the labeled images. Second, the construction of neural networks follows MobileNet, we should incorporate the learned feature from hash function into binary codes, and add a latent layer with unit layer to the top layer. Each layer is utilized by a batch-norm layer and a nonlinear layer with the fully-connected classification layer. The average pooling layer is used to reduce the spatial resolution before the binary codes.

N images is denoted as $\{I_n\}_{n=1}^{N}$; vectors denoted as y_n^{MN} are associated with labels, in which y_n denote an entry with a value of 1 of the image classification belongs to the corresponding class. The mapping relationship is presented as $F : I \rightarrow \{0,1\}^{M\times K}$, where the $k-$ bit binary codes $B = b_n \in \{0,1\}^{M\times K}$ is preserved as semantic similarity among the crowdsourced image database.

Algorithm 3: BOW Based Deep Hash Method Detection

Input: a object set $O = \{O_1, O_2, \ldots, O_i, \ldots\}$
Output: $W : [V1, V2, \ldots, Vn]$
// V denotes visual word
foreach $i \leftarrow 1$ *to* $|O|$ **do**
  foreach $j \leftarrow 1$ *to* $|O_i| - 1$ **do**
    foreach $l \leftarrow j + 1$ *to* $|O_i|$ **do**
      if $f \in image[j]$ *and* $f' \in image[l]$
      satisfy eq.(7, 8) **then if**
      $f \in image[j]$ *and* $f' \in image[l]$
      satisfy *eq.*(4, 10)
      anddistance *satisfyeq.*(11) **then** {
        $w_p \leftarrow 1$;
        $w_q \leftarrow 1$;
      }
      /* p corresponds to f ID;
      q corresponds to f' ID */ ;
    end
  end
end

V. SPATIAL DIVISION

Spatial division is one of the most important element to consider in Fig. 4. The most important issues are focus

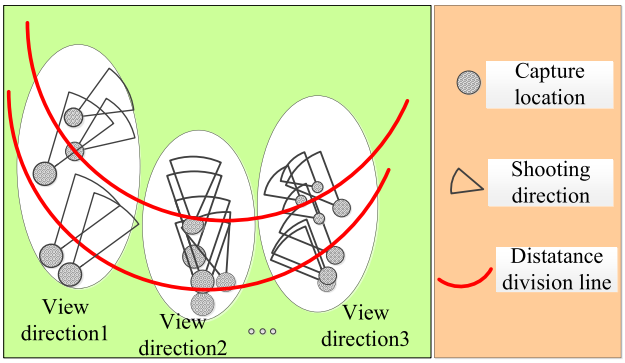

FIGURE 4. Examples of spatial division in the crowdsourced images.

on coordinate location, shooting direction and distances. In Fig. 4, the red line divides the distance, which is represented as short, middle and long distance. The distance scale D_s is also clustered by the adaptive situation with respected to the specifical object by the Eq. 11.

$$D\_s = \left\lceil \frac{|Dis_{max} - Dis_{min}|}{Dis_{max} + Dis_{min}} \right\rceil \tag{11}$$

D_s is calculated by Dis_{max} and Dis_{min} in Eq.11, in which the distance is determined by Dis_{max} and Dis_{min}. In Fig. 4, the coordinate location is represented as GPS and the shooting direction is calculated by the sensory data. The overall distance is segmented by the distance between the object and the sensory camera.

VI. VOTING FOR THE BEST MATCHING IMAGES

Locality-Sensitive Hashing (LSH) [20] is built for feature matching, which is indexed by Fast Library for Approximate Nearest Neighbors (FLANN) [20]. Suppose Features set $S = [s_1, s_2, \ldots, s_f]$ is extracted from one query image, the top t matching features are extracted from every feature in S. With the LSH index, every query image is associated with the best matching feature ID matrix in S.

$$S = \begin{pmatrix} s_{11} & s_{12} & \cdots & s_{1t} \\ s_{21} & s_{22} & \cdots & s_{2t} \\ \vdots & \vdots & \ddots & \vdots \\ s_{f1} & s_{f2} & \cdots & s_{ft} \end{pmatrix}$$

In the matrix, s_{ft} denotes the best matching feature ID for every element in the matrix S. Specifically, s_{ft} means that the s-th query feature's top t-th best matching feature ID is f_{st}. With respect to the best matching image, we can obtain the feature vote results in Eq. (12).

$$V: \quad [s_1 * w_1, s_2 * w_2, \ldots] \tag{12}$$

After obtaining the feature vote vector V, it is essential to search for the best matching location images. It is notable that the features corresponding to the identical image are labeled with the image ID at the stage of extracting the SURF features from image database. In other words, we are capable of obtaining the corresponding mapping matrix which records the relationship of feature ID and image ID. Most importantly, the image vote vector can be obtained through feature vote vector with Eq. (13) by the relationship mentioned above. In Eq. (13), the vote of the i-th image is denoted by V_i.

$$I: \quad [V_1, V_2, \ldots V_i, \ldots] \tag{13}$$

The best matching images are obtained through returning the image searching result with the best matching numbers ranked in top N. It is obvious that RCCF based VHB scheme is suitable for crowdsourced image searching, due to the rich informative advantage mentioned above. Whereas, Oxford Building Dataset does not have the advantage of the crowdsourced images. For an example, images in Oxford Building Dataset are quite different from each other in appearance. As a result, images taken in the building may share few common features with the images taken outside the identical building.

VII. EXPERIMENT

In this section, we conduct the experiments and evaluate the results on the crowdsourced image database. Then, we compare the experimental performances with respect to RCCF scheme and related methods. To demonstrate the effectiveness, we compare the RCCF based VHB scheme with other similar schemes, in contrast with existing approaches, e.g., VHB and CCF based CHB.

A. DATABASE AND IMPLEMENTATION

Extensive experiments are conducted on the BUPT crowdsourced image dataset [5], which covers typical buildings, for example, teaching rooms, gymnasium, and student dormitories on the campus. In the BUPT dataset, the total dataset [5] is consisted of 8,062 images crowdsourced from 162 objects which contain library, buildings and dormitories. Among the images, 1,620 images along with the rich sensory data are randomly selected as testing dataset. In addition, the rest 6,442 images are saved as training set, which has rich sensory data to obtain shooting angle and distance scale.

The deep hash net structure is conducted with tensorflow platform, which is initialized by the parameters with the parameters of a MobileNet trained on 14 million images of the ImageNet dataset. The hash layer parameters is initialized with the learning rate as 0.01, which is decreased to 1/15 of the previous value. The training process is conducted with 30,000 iterations with a mini-batch if 32 images to minimize the classification error.

Meanwhile, the RCCF based VHB scheme is conducted with the environment of 3.2 GHz CPU and 8 GB memory. The OpenCV library is utilized to extract 968,546 SURF features from training set while 73,954 features are non-CCF features. On average, 130 features are extracted from every image query. Moreover, the spectral hash function is trained in Matlab 2011 to obtain the transition parameter matrix. FLANN is used to implement immediately best features in binary code matching.

B. BASELINES

We compare RCCF based VHB with several similar methods, including VHB method, BOW method, CCF based VHB scheme. Because of the extremely high computational cost with BOW method, the visual dictionary is trained with 8,623 visual words, in which 8,623 BoW histogram is compressed into 1500 dimensions by hierarchical K-means.

Both Precision@N and MAP@N are utilized to evaluate the performance of our scheme. The metrics are widely used in the state-of-the-art location recognition systems [2], [3], [10]. Precision@N indicates the top 1 image candidate with the top N best matching features as follows:

$$Precision@N = \frac{1}{N_q}\sum_{i=1}^{N_q}(P_i(\sum_{f=1}^{N_f}(\sum_{v=1}^{N}V(v))) \tag{14}$$

In Eq. 14, N_q is the number of query image; P_i is the precision of the ith query image; N_f is the number of the features in the ith query image; $V(v)$ is the vote of vth best matching feature from the training database.

Mean average precision at N ($MAP@N$) [2], [3] is used to evaluate the proportion of correct matches in the top N image candidates, revealing the position-sensitive ranking precision of the queries as follows:

$$MAP@N = \frac{1}{N_q}\sum_{i=1}^{N_q}(\frac{\sum_{r=1}^{N}P(r)rel(r)}{N}) \tag{15}$$

C. EXPERIMENTAL COMPARISON

Extensive experiments are conducted on BUPT Dataset [5]. With respect to the best matching features N in Eq.14, it is set from 1 to 10. In Fig. 5, we compare the performance of RCCF based VHB scheme with similar methods, including CCF based VHB scheme and VHB approach. From Fig. 5, it is observed that RCCF based scheme obtains higher precision than similar approaches. Because of utilizing the richer information mentioned above, RCCF based scheme obtains higher precision, including various kinds of sensory data. It is notable that small number of best matching features may cause the incorrect results; whereas large number of best matching features may also cause some incorrect results. It is explained that when the number of best matching features is set as 5, the overall precision obtains the optimal performance. It is concluded that RCCF based method effectively improves the searching accuracy through utilizing richer sensory data, excluding the disturbances in Section I.

In the experiments, we also consider the position-sensitive ranking precision, i.e., the Eq. 15 indicates the performance, when N varies also from 1 to 10; $P(r)$ is the precision at the cut-off rank of r; $rel(r)$ is binary function on whether r is a correct match. In Fig 7, it is shown that RCCF based VHB method obtains the best matching results, outperforming other similar methods, for example, CrowdLR, VHB and CCF based VHB method. It is worth to mentioning that in this experiment CrowdLR system adopts BOW image searching method, with the codebook containing approximate

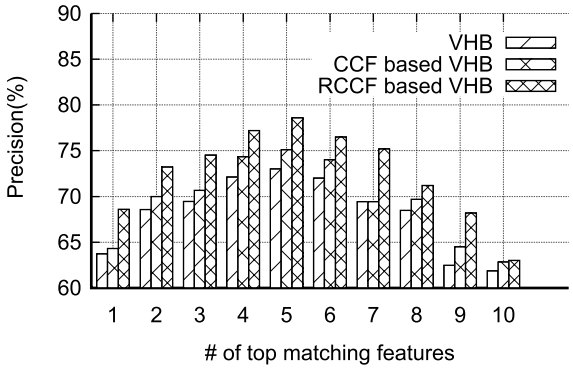

FIGURE 5. Precision @N returns a best matching image with top N features which vary from 1 to 10.

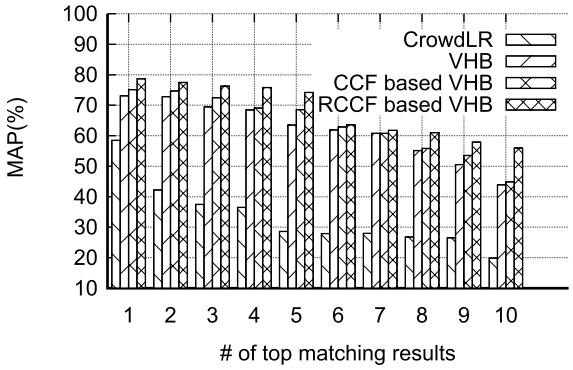

FIGURE 6. MAP@N return the top N images which vary from 1 to 10.

50,000 visual words. Furthermore, we set the number of top matching features as 5 in the experiment, similar experimental results are obtained if adjusting it. From the results, it is shown that RCCF based VHB method obtains around 5% higher precision with the position-sensitive comparison if returning multiple results with ranking order. It is also confirmed that our RCCF based VHB scheme decreases the negative impacts of the disturbances in the crowdsourced image database.

We conduct the experiments with deep feature extraction with visual search, on the crowdsourced location dataset [7]. The mobileNet architecture is utilized to solve the efficient search on the mobile devices, using the parameters of a trained net on the 14 millions images [7]. The training rate is initialized as 0.01, terminating at iterations 30,000 iterations. To train the network parameters, the back propagation rate is set with a mini-batch size to minimize the classification.

In Table 1, we evaluate the precision performance by MAP with respect to rank metric, considering Hamming distances from query images. In this experiment, we select the *top K* images from the ranked list as the retrieval results. We compare the MAP performance with different hash code bits with different location recognition methods. From Table 1, it is observed that the precision of deep hash method greatly outperforms other similar methods, which demonstrate the deep learning method for binary code training.

TABLE 1. MAP with different hash bit.

Method	MAP %			
	16-bit	32-bit	64-bit	128-bit
VHB	-	-	19.36	21.98
ALexNet	55.69	59.23	60.32	68.98
MobileNet	53.60	59.99	62.23	70.82
CCF based VHB	59.96	76.54	87.15	76.54
RCCF based VHB	**62.68**	**77.65**	**88.45**	**79.32**

To demonstrate the benefits of RCCF based method, we also conduct the experiments on BoW, utilizing the compressed histogram into 100, 000 and 1,000, 000 dimensions. In addition, we use the codebook which is consisted of 9,945 visual words to test the related methods, adopting OpenCV library to extract 661,522 SURF features from the training set. From Table 2, it is shown that RCCF based method outperforms other similar methods, with codebook size 100,000 and 1,000,000 respectively. Adopting rich sensory data, our method obtains the optimal precision with different codebook length.

TABLE 2. MAP with BOW method.

Method	MAP % (codebook)				
	BOW	ALexNet	MobileNet	CCF based BOW	**Ours**
100,000	55.69	59.23	60.32	68.98	**70.65**
1,000,000	53.60	59.99	62.23	70.82	**71.32**

Spatial division is also an important issue, which focus on the distance of Dis_{max} and Dis_{min}. We conduct the spatial division experiment, which varies distance scale represented as short, middle and long. For fair comparison, the returning hash bit is equally set as 64 bit and the best matching image number is set as 6. It is worth mentioning that the distance scale is determined the specific distance of one object, i.e., different objects share different distance scale due to the different shooting angle and distances.

TABLE 3. MAP with distance scale.

Method	MAP % (Distance Scale)		
	short distance	middle distance	long distance
VHB	-	19.62	22.18
ALexNet	59.23	65.24	66.83
MobileNet	62.92	62.36	74.22
CCF based VHB	75.47	88.53	78.46
RCCF based VHB	**76.37**	**89.01**	**78.23**

We conduct the last experiment to testify the record of shooting angle and mean average precision, which is also an important issue of mobile location recognition. As shown in Fig. 7, it is observed that the RCCF based VHB method outperforms other similar methods. It compares the MAP of location recognition with various degree of shooting angle. It is shown that the MAP increases with the increase of degree, because too small degree causes too much deviation of the main facade of buildings. However, too large degree also causes too much deviation. As a result, Fig 7 shows that too large degree also causes the decrease the MAP. It is also confirmed that

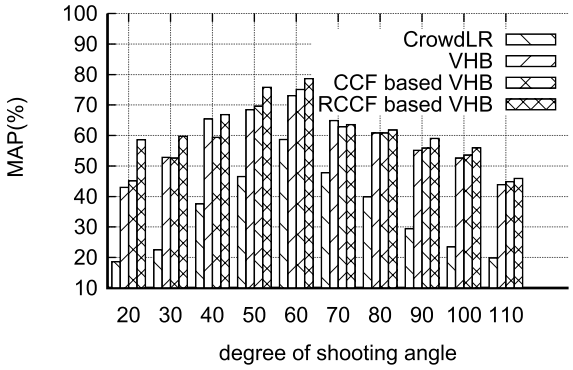

FIGURE 7. MAP@*N* vs. shooting angle degree which varies from 20 to 110.

VIII. CONCLUSION AND FUTURE WORK

In this paper, we propose an RCCF based VHB scheme to decrease negative effects of various disturbances. Furthermore, we utilize the deep hash method to improve various performance metrics. Extensive experiments show that the rich information of crowdsourced images can be better leveraged to improve the accuracy performance. In the future, we will fuse multiple types of sensing data (e.g., GPS location, distance, angle, etc.) to further improve various performance metrics. We will also further explore promising benefits of our scheme through more evaluations on other datasets.

REFERENCES

[1] V. Raychoudhury, S. Shrivastav, S. S. Sandha, and J. Cao, "CROWD-PAN-360: Crowdsourcing based context-aware panoramic map generation for smartphone users," *IEEE Trans. Parallel Distrib. Syst.*, vol. 26, no. 8, pp. 2208–2219, Aug. 2015.

[2] X. M. Qian, Y. Xue, X. Y. Yang, Y. Y. Tang, X. Z. Hou, and T. Mei, "Landmark summarization with diverse viewpoints," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 25, no. 11, pp. 1857–1969, Nov. 2014.

[3] R. R. Ji, L. Y. Duan, X. Y. Yang, Y. Y. Tang, X. Z. Hou, and T. Mei, "Landmark summarization with diverse viewpoints," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 25, no. 3, pp. 1857–1969, Nov. 2014.

[4] X. W. Li, C. C. Wu, X. Zach, S. Lazebnik, and J. M. Frahm, "Modeling and recognition of landmark image collections using iconic scene graphs," in *Proc. Eur. Conf. Comput. Vis.*, Oct. 2008, pp. 427–440.

[5] H. Wang, D. Zhao, H. Ma, and L. Ding, "MB-GVNS: Memetic based bidirectional general variable neighborhood search for time-sensitive task allocation in mobile crowd sensing," *IEEE Trans. Veh. Technol.*, vol. 69, no. 2, pp. 2219–2229, Feb. 2020.

[6] D. Zhao, H. Wang, H. Ma, H. Xu, L. Liu, and P. Zhang, "CrowdOLR: Toward object location recognition with crowdsourced fingerprints using smartphones," *IEEE Trans. Human-Mach. Syst.*, vol. 47, no. 6, pp. 1005–1016, Dec. 2017.

[7] H. Wang, D. Zhao, and H. Ma, "Informative image selection for crowdsourcing-based mobile location recognition," *Multimedia Syst.*, vol. 25, no. 5, pp. 513–523, Oct. 2019.

[8] H. Wang, D. Zhao, H. Ma, H. Xu, and X. Hou, "Crowdsourcing based mobile location recognition with richer fingerprints from smartphone sensors," in *Proc. IEEE 21st Int. Conf. Parallel Distrib. Syst. (ICPADS)*, Dec. 2015, pp. 156–163.

[9] H. Wang, D. Zhao, H. D. Ma, and H. Y. Xu, "SSFS: A space-saliency fingerprint selection framework for crowdsourcing based mobile location recognition," in *Proc. Pacific–Rim Conf. Multimedia*, Sep. 2016, pp. 650–659.

[10] X. Yang, X. Qian, and Y. Xue, "Scalable mobile image retrieval by exploring contextual saliency," *IEEE Trans. Image Process.*, vol. 24, no. 6, pp. 1709–1721, Jun. 2015.
[11] W. Liu, T. Mei, and Y. Zhang, "Instant mobile video search with layered audio-video indexing and progressive transmission," *IEEE Trans. Multimedia*, vol. 16, no. 8, pp. 2242–2255, Dec. 2014.
[12] S. Rudinac, M. Larson, and A. Hanjalic, "Learning crowdsourced user preferences for visual summarization of image collections," *IEEE Trans. Multimedia*, vol. 15, no. 6, pp. 1231–1243, Oct. 2013.
[13] L. von Ahn and L. Dabbish, "Labeling images with a computer game," in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, Apr. 2004, pp. 319–326.
[14] H. Chen, B. Guo, Z. Yu, and L. Chen, "CrowdPic: A multi-coverage picture collection framework for mobile crowd photographing," in *Proc. IEEE 12th Int. Conf. Ubiquitous Intell. Comput. IEEE 12th Int. Conf. Autonomic Trusted Comput. IEEE 15th Int. Conf. Scalable Comput. Commun. Associated Workshops (UIC-ATC-ScalCom)*, Aug. 2015, pp. 68–76.
[15] Y. Wang, W. J. Hu, Y. B. Wu, and G. H. Cao, "SmartPhoto: A resource-aware crowdsourcing approach for image sensing with smartphones," in *Proc. Int. Symp. Mobile Ad Hoc Netw. Comput.*, Jun. 2014, pp. 113–122.
[16] H. Bay, T. Tuytelaars, and L. V. Gool, "SURF: Speeded up robust features," in *Proc. 9th Eur. Conf. Comput. Vis.*, Graz, Austria, May 2006, pp. 404–417.
[17] J. F. He, J. Y. Feng, X. L. Liu, T. Cheng, T. H. Lin, H. J. Chung, and S. F. Chang, "Mobile product search with bag of hash bits and boundary reranking," in *Proc. IEEE Conf. CVPR*, Jun. 2012, pp. 3005–3012.
[18] Y. Weiss, A. Torralba, and R. Fergus, "Spectral hashing," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, pp. 1753–1760, 2009.
[19] W. Liu, T. Mei, Y. Zhang, J. Li, and S. Li, "Listen, look, and gotcha: Instant video search with mobile phones by layered audio-video indexing," in *Proc. 21st ACM Int. Conf. Multimedia (MM)*, Oct. 2013, pp. 887–896.
[20] I. Piotr and M. Rajeev, "Approximate nearest neighbor: Towards removing the curse of dimensionality," *Theory Comput.*, vol. 25, no. 5, pp. 604–613, Jan. 2012.
[21] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Annu. Conf. Adv. Neural Inf. Process. Syst.*, Lake Tahoe, NV, USA, Dec. 2012, pp. 1106–1114.

HAO WANG received the B.S. and M.S. degrees from the School of Computer Science, Henan University, China, in 2006 and 2009, respectively, and the Ph.D. degree from the School of Computer Science, Beijing University of Posts and Telecommunications, Beijing, China. He is currently a Faculty Member with the Nanyang Institute of Technology, Henan, China. His research interests include crowdsourcing and multimedia computing.

YUGUI WANG received the B.S. degree from Henan University, China, in 2006, and the M.S. degree from Xidian University, China, in 2010. He is a Lecture in Nanjing Polytechnic Institute, Nanjing, China. His research interests include image processing and sensory data processing.

RUI CUI received the B.S. degree from Northeast Normal University, in 2004, and the M.S. degree from Nanyang Normal University, China, in 2015. She is currently a Lecturer with Nanyang Normal University, Henan, China. Her research interests include image processing and big data.

YIBO HAN received the M.S. and Ph.D. degrees from the China University of Mining and Technology, Beijing, in 2009 and 2015, respectively. He is currently an Associated Professor with the Nanyang Institute of Technology, Henan, China. His research interests include the IoT and sensor.

CHAOHUA YAN recieved the B.S. degree from Nanyang Institute of Technology in 2010. He is a Lecture in Nanyang Institute of Technology. His research interests include information processing and data analysis.

MENGHAN NIU is currently pursuing the M.S. degree with the Nanyang Institute of Technology, Henan, China. Her research interests include image processing and machine learning.