Image Location Inference by Multisaliency Enhancement

Xueming Qian, *Member, IEEE*, Huan Wang, Yisi Zhao, Xingsong Hou, Richang Hong, *Member, IEEE*, Meng Wang, *Member, IEEE*, and Yuan Yan Tang, *Fellow, IEEE*

***Abstract*—Locations of images have been widely used in many application scenarios for large geotagged image corpora. As to images that are not geographically tagged, we estimate their locations with the help of the large geotagged image set by content-based image retrieval. Bag-of-words image representation has been utilized widely. However, the individual visual word-based image retrieval approach is not effective in expressing the salient relationships of image region. In this paper, we present an image location estimation approach by multisaliency enhancement. We first extract region-of-interests (ROIs) by mean-shift clustering on the visual words and salient map of the image based on which we further determine the importance of the ROI. Then, we describe each ROI by the spatial descriptors of visual words. Finally, region-based visual phrases are generated to further enhance the saliency in image location estimation. Experiments show the effectiveness of our proposed approach.**

***Index Terms*—Location estimation, region of interest (ROI), visual phrase, salient map, salient region, spatial constraint.**

I. Introduction

TODAY, social media provide a platform for us to share messages with our friends. Especially with the development of smartphone, it is very convenient for us to access internet and take photos at any time and any place. Recently, most of the cameras can embed geo-graphic locations in the photos. Photos taken at place-of-interests (POIs) are extremely popular for users to share. The social media websites such as Flickr, Picasa gather billions of photos shared by world-wide users. Thus from the large scale geo-tagged images, we can carry out image-location estimation for the images without GPS information [4], [35], [68]. Based on the geo-graphical location of images, we can infer the locations of user, which is useful for location based services recommendation [67], user preferred POIs recommendation [28], and user travel footprint mining and management [42].

The image location estimation is actually a content based image retrieval or recognition [4], [35]. In this paper, our task is to estimate the location of an input image by mining image content. As to images which are not geographically tagged, we estimate their locations with the help of the large geo-tagged image set. State-of-the-art large scale image retrieval systems have relied on the BoW model [1]–[9], [14]–[16], [34]–[37], [58] and local descriptors. And the idea of hierarchical vocabulary tree [34] accelerates the speed of clustering and quantizing for large scale image retrieval. Traditionally, a visual vocabulary is trained by clustering a large number of local feature descriptors, such as SIFT, and SURF [50]. The exemplar descriptor of each cluster is called a visual word, which can be indexed by an integer in fast image retrieval [58].

Experimental results of existing work show that the commonly generated visual words are still not so expressive. In this paper, we further exploit saliency from the bag-of-words to improve the image location estimation performance. Quantization limits the discriminative power and ignores geometric relationships among visual words [8], [19], [58]. Spatial verification enforces geometric consistent constraint on visual words that query and dataset images [8], [19], [35], [58]. Spatial information of visual words should be exploited for better image retrieval performance. Wu *et al.* [36] employ the detector of Maximally Stable Extremal Regions (MSER) to bundle SIFT features into groups.

Similarly, we present a ROI based location estimation approach by considering that the distribution of an image's visual words directly reflects the distribution of the image's main content [35]. We mine the salient regions for location estimation and exploit spatial information for each region's visual words. Firstly, we divide an image's visual words into groups by Mean-shift clustering [35]. In this process, the coordinates of BoWs are utilized. A ROI is composed of visual words in the corresponding cluster. Then within each ROI, spatial coding for visual words is conducted. Because the bundled feature based methods employ group feature matching instead of single feature matching, the ROI based location estimation approach has more discriminative performance than the methods using the single SIFT feature.

Based on our previous work [35], we further propose a multi-saliency enhancement based image location estimation

Manuscript received August 14, 2015; revised February 18, 2016 and August 2, 2016; accepted September 9, 2016. Date of publication December 9, 2016; date of current version March 15, 2017. This work was supported in part by the Program 973 under Grant 2012CB316400, in part by NSFC under Grant 60903121, Grant 61173109, and Grant 61332018, in part by Microsoft Research Asia, and in part by the Fundamental Research Funds for the Central Universities. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. K. Selcuk Candan. *(Corresponding author: Xueming Qian.)*

X. Qian, H. Wang, Y. Zhao, and X. Hou are with the Smiles Laboratory, School of Electronics and Information Engineering, Xi'an Jiaotong University, Xi'an 710049, China (e-mail: qianxm@mail.xjtu.edu.cn; email_wanghuan@163.com; zys2012@stu.xjtu.edu.cn; houxs@mail.xjtu.edu.cn).

R. Hong and M. Wang are with the Hefei University of Technology, Hefei 230009, China (e-mail: hongrc.hfut@gmail.com; eric.mengwang@gmail.com).

Y. Y. Tang with the Faculty of Science and Technology, Macau University, Macau, China (e-mail: yytang@umac.edu.cn).

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

Digital Object Identifier 10.1109/TMM.2016.2638207

approach. The saliency is explored mainly from the following three aspects: salient region extraction, salient map extraction and salient region representation. The main contributions of this paper can be summarized as following:

1) We propose a multi-saliency enhancement based image location estimation approach. In the proposed approach, we enhance the saliency of the input image and dataset images by considering the visual word distribution information, visual saliency and the visual phrase representation in a unified framework.
2) We propose a multi-region ranking based similarity measurement approach to differentiate the importance value of a ROI by taking the visual saliency information into account.
3) We propose a region based visual phrases representation approach, which generates the spatial descriptor for each visual word in salient region. The region based visual phrase captures the relative spatial distribution of the visual words in the region, it is effective in geometric consistency checking.
4) We build a fast indexing structure, which records the saliency information. From the fast indexing structure we can determine the location of the input image quickly. The fast indexing approach require less computational costs compared with the without indexing approach. Comparing with the-state-of-the-art approaches, it achieves better performance with a bit computation cost.

The rest of the paper is organized as follows: in Section II, related works are reviewed. In Section III, we provide the system overview and the detailed approach. In Section IV the experiments and discussions are given. In Section V, the conclusions are drawn.

II. Related Work

In this paper, we propose an image taken place estimation approach by multi-saliency enhancement. We give a comprehensive overview of the related work on following two aspects: 1) visual feature image location estimation, 2) image retrieval by saliency enhancement.

A. Visual Feature-Based Image Location Estimation

In recent years, visual feature based image location estimation approaches have been paid much attention [2]–[9]. Most of the existing approaches utilize the visual information to carry out content based image search. In the existing works, the BoW models are well developed. For example, Li *et al.* utilize multi-class SVM classifiers for large scale image location estimation [2]. They also fuse the textual information to improve the visual based location estimation performances. Quack *et al.* propose to utilize local feature matching to carry out image location estimation [23]. To speed up the estimation process, user interaction is required to place the input images to a rough geographic area.

Li *et al.* propose a global feature clustering and local feature matching based fast image location estimation approach [4]. They adopt the inverted file structure (IFS) and use representative images for each GPS location to guarantee the estimation speed and accuracy. They further propose a salient feature mining based image location estimation approach [38]. They first mine the salient region of the input image by exploring its relation with k nearest neighboring image groups, and then select salient features by considering their relationship with the neighbor image groups.

Donoser *et al.* propose of match interest points detected in the query image to a sparse 3D point cloud [3]. They project features to fern-specific embedding spaces to improve the performance. Then, the obtained correspondences are used to estimate camera pose.

Zhou *et al.* proposed a spatial coding based image retrieval approach [8]. It encodes the relative positions between each pair of features in an image. Zhang *et al.* propose a spatial coding based image retrieval approach by building contextual visual vocabulary [31].

And some new technologies continuously emerge, such as domain-adaptive global feature descriptor [41], re-ranking schemes for dataset images [32]. Some trivial branches of vocabulary tree are pruned to decrease the dimension of BoW. Sparse coding compresses the original BoW histogram of query into several coefficients [29], [30], [39]. Thus the high dimensional BoW histogram is projected into a low dimensional vector via transformation matrix or dictionary. Moreover, the database can be constructed with a 3D model. [6], [13], [17] and [20] are related to GPS location estimation using constructed 3-D models from large scale geo-tagged photos. Liu *et al.* proposed an approach which is capable of providing a complete set of parameters about the geo-information—including the actual locations of both the mobile user and the scene along with the viewing direction [6]. They first perform joint geo-visual clustering to generate scene clusters and represent each scene by a 3D model. The 3D scene models are then indexed using a visual vocabulary tree structure. [17] carries out the image retrieval by generating 3-D models and translating the query image into a 3-D pattern. Park el al. proposed a view direction determination by utilizing Google Street View and Google Earth satellite [20].

B. Image Retrieval by Saliency Enhancement

Saliency detection is a very popular topic in object detection [74], scene recognition [69], [72], [76] and image retrieval [7], [19], [25], [27], [35], [75]–[79], and other image based applications [69]–[72]. The traditional saliency detection algorithms, utilize a variety of hand-crafted features, such as the texture descriptors, and color histogram etc. from a single image [51]. While recently, the co-saliency detection approaches are very popular by utilizing multiple same style image that share similar objects [53], [54]. The co-salient objects in a set of similar image group have similar low-level features, such as, textures, colors and some local feature descriptors. Cheng *et al.* [52] propose a histogram-based contrast method to define salient value for each pixel. They also propose a saliency cut based salient region extraction approach, which uses the salient map to detect salient object and extract salient region. Li and Ngan propose to utilize conventional saliency detection approaches to determine the single-image saliency and to obtain the multi-image saliency [53]. Han *et al.* propose a deep learning based co-saliency detection approach [54]. They look deep to transfer

high-level representations by the convolutional neural network (CNN) with adaptive layers which could enhance the properties of the co-salient objects. Fu *et al.* propose cluster-based co-saliency detection approach [56]. They learn global correspondence between the multiple images during clustering. They fuse contrast and spatial clues to measure the cluster saliency. In [57], they rebuild the images and provide the reconstruction error regarded as a negative correlational value in co-saliency measurement.

However, as far as we know, multi-saliency enhancement is not well utilized in image retrieval. Usually, researchers try to extract saliency from handicraft features [4], [19], [31], [35], [55] and deep learning features [60]–[65] to improve retrieval performances.

The SIFT feature and BoW model have been manifested in image retrieval. However, the BoW models have some deficiency. For example, owing to the quantization loss, the visual word is not discriminative enough. Thus, many improved approaches are proposed to enhance the discrimination of BoW for image retrieval, e.g. visual synonyms [7], [19], [34], [40], [41], [58], embeding geometry constraint [1], [8], [19], [37], [58], [59], [68], and spatial verification [19], [31], [35], [58].

Gavves *et al.* define visual synonyms as pairs of independent visual words that could be mapped to each other in similar images via a trained homographic matrix [7]. Spatial information [19], [26], [35], [58] can be utilized to enhance the discriminative power of single visual word.

Chum *et al.* explore salient information from multiple relevant images founded from the initial searching steps to improve the performance of image retrieval. Yang *et al.* proposed to explore the contextual saliency information that mined from multiple queries to improve image retrieval performances [19], [55]. They further rank the saliency for each visual word or visual phase to enhance its robustness in image retrieval scalable mobile image retrieval with geometric consistency checking [35], [58], [59]. They show that the contextual saliency can not only improve the performance, but also reduce the quantity of the data that needs transmitted from mobile end to cloud/server end.

Recently, deep learning based approaches are very popular object detection and recognition. The approach can learn salient information from images deep image features, which should be helpful for mining salient information for retrieval. Krizhevsky *et al.* use the feature vectors from the last layer of CNN in image retrieval and demonstrated outstanding performance on ImageNet [60]. Many researches show that directly utilizing the deep feature rather than the low-level visual feature, better performances can be achieved [61], [62]. Zhang *et al.* propose a deep learning based hashing algorithm form image retrieval [63]. They organize the training images into a batch of triplet samples, each sample containing two images with the same label and one with a different label. The deep convolutional neural network is utilized to train the model to generate efficient descriptors for the salient feature detection.

However, the existing saliency detection approaches based on CNN are far more complex. In this paper, we further explore the saliency in a ROI based on handicraft features to improve image location estimation performance.

Fig. 1. Block diagram of the location estimation system.

III. The Proposed Approach

The main flowchart of the proposed approach is shown in Fig. 1. It consists of online and offline parts. The online part mainly consists of the following four steps: 1) ROI generation, 2) ROI saliency determination, 3) ROI description, and 4) similarity measurement. The major parts of the offline system and online system are the same. The only difference is that we need to build salient index for the dataset image with GPS information.

A. ROI Generation

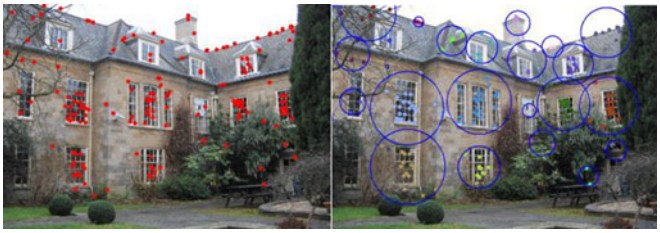
In this paper, we aim at representing image by the salient region rather than using single visual words. The main motivation of our approach is that the salient region information is far more robust than BoW [35]. The bundled feature based methods employ group feature matching instead of single feature matching. Thus, in this section, we divide the image into ROIs to bundle visual words.

For an image, we cluster the coordinates of its useful visual words by mean-shift clustering [28], [35]. Usually, each SIFT point has a 128-D descriptor vector and a 4-dimensional DoG key-point detector vector (x, y, scale, and orientation). Here the coordinates of visual words are utilized. Let $v = \{(x_i, y_i)\}_{i=1}^{h}$ denote the locations of the h -th SIFT points. To $\forall v$, mean-shift is defined as follows:

$$\begin{cases} M_b(v) = \dfrac{1}{N_o} \sum_{v_i \in S_b(v)} v_i - v \\ S_b(v) = \left\{ z : (z-v)^T (z-v) \le b^2 \right\} \end{cases} \tag{1}$$

where $S_b(v)$ is the region whose radius is b and whose centroid is v. N_o is the number of SIFT points falling within the region $S_b(v)$. z is the visual words falling within $S_b(v)$.

After mean-shift based visual words clustering, we obtain several clusters (i.e. ROIs) and their corresponding centers. A cluster is represented by a visual word group (i.e.VWG) [35], which is composed of the visual words in its cluster. For an image, we assume that there are L ROIs generated, and we denote them as $G_l, l = 1, 2, \ldots, L$. We only keep the SIFT feature points with high tf-idf values as shown in Fig. 2(a). Some, un-discriminative SIFT feature points are removed before clustering. The corresponding regions after mean-shift clustering each ROI are shown in Fig. 2(b). We mark a unique color for the visual words. Totally there are 24 ROIs. We represent an image by ROIs, each of ROI is composed of a set of visual words. By comparing Fig. 2(a) and Fig. 2(b), we find that, the number of ROIs is far less than that of SIFT points.



(a) (b)

Fig. 2. ROI representation for an image.

B. ROI Saliency Determination

In this section, we carry out ROI saliency determination and ranking for our image location estimation. The saliency of a ROI can be determined from both the salient map of an image and ROI characteristic.

1) Saliency From Salient Map: In this paper, we adopt the approach [51] to determine the salient map of an image. We first compute the multi-scale intensity, color and orientation features of the image to get the final salient map. Let $S(x, y)$ denote the salient map of an image. Based on the obtained ROIs G_l, we determine the importance of each ROI by taking into the salient map into consideration. A ROI being salient or not is measured from its corresponding saliency map as follows:

$$p(l) = \frac{1}{|G_l|} \sum\nolimits_{(x,y)\in G_l} S(x, y) \tag{2}$$

where $|G_l|$ denote the pixel number in the ROI G_l. If the major pixels in a ROI are with higher saliency, then this region will have high importance in location estimation.

2) Saliency From ROI: As different ROIs have different sizes and at different locations, their contributions for retrieval should be different. Generally speaking, the large the relative size of the ROI, the more visual words it contains, and the more important that this region in image location estimation. Let $r(l)$ be the relative size of the ROI, which is determined as follows:

$$r(l) = \frac{n_l}{N}; l = 1, \ldots, L \tag{3}$$

where n_l is the number of visual words in ROI G_l and N is the total number of visual words in the image.

Finally, we fuse the saliency from both salient map and ROI in determining the final weight of a ROI as follows:

$$\text{weight}_l = p(l) \times r(l); l = 1, \ldots, L. \tag{4}$$

Based on the weights, we can rank the importance of each ROI. The larger the weight, the higher the contribution of the ROI to the image location estimation.

C. ROI Description

Assuming that a ROI G_l has n visual words which are denoted by $\{w_1, w_2, \ldots, w_n\}$, our position descriptor includes the relative area (RA), and the relative distance (RD). Our ROI description approach includes enhanced saliency by visual phrase extraction and position descriptor for each visual word. We first extract visual phrase and then generate position descriptor for a visual word to describe its distribution in the corresponding ROI. More detailed illustration for ROI representation can be found in [35].

1) Visual Phrase Representation: Firstly, we calculate the spatial distance of two visual words by their SIFT feature descriptors. Let (x_i, y_i) and (x_j, y_j) denote the coordinates of visual words w_i andw_j, their spatial distance $d(w_i, w_j)$ is calculated by the Euclidean distance.

Secondly, any two visual words in a salient region can be viewed as a visual phrase if their distance is smaller than a threshold thr. In this way, a ROI is represented by a set of visual phrases, denoted by P, with each element as follows:

$$P = \{vp_i\}_{i=1}^{Q} = \left\{\left(W_i^1, W_i^2\right)\right\}_{i=1}^{Q}, d\left(W_i^1, W_i^2\right) \le thr \tag{5}$$

where vp_i is the i-th visual phrase, W_i^1 and W_i^2 are the corresponding two visual words, Q is the total number of visual phrases in the ROI.

2) RA Representation: We set the center of a cluster as the center of the corresponding ROI. For a visual word w_i, we record its relative area RA to the center of the ROI. For each visual word w_i in G_l, we record its relative spatial position against the origin. When the visual word is a bottom-right word, we define that its RA is [0 0 0 1]. Thus the RA of a visual word is a 4 bit descriptor.

3) RD Representation: We calculate the relative distance RD between the visual word and the center of the ROI. That is to say we want to know whether the distance between visual word and its corresponding center of ROI is large relatively. For a visual word, if its distance to the center is less than the average distance, the visual word is near the center, otherwise, we think that the word is relatively far away from the center. Accordingly RD is a 1-bit spatial descriptor, which reflects the visual word's distance is relatively far from the center of the region or not. For a visual word w_i its position descriptor combines both relative area RA_i and relative distance RD_i.

D. Image Indexing

For the input query image the final ROI description is shown in Fig. 3(a). We record the corresponding weight, the visual phrases, and the corresponding spatial descriptors for each of the two visual words in a ROI.

For fast location estimation, we build inverted index for each visual phrase and its corresponding images as shown in Fig. 3(b). For each image, we record the corresponding ROI, weight and the spatial descriptors for the two visual words. The GPS location of image $\#I$ which is denoted by $Label_I$ is all recorded in another image-location inverted file list. The position descriptor of visual word w_x including a four bits RA and one bit RD is also recorded.

E. ROI-Based Similarity Measurement

We formulate the image retrieval as a voting problem. Each visual phrase and the weight of each ROI in the query image votes on its matched images. For each visual phrase occurred in the query image $\#q$, we use the obtained inverted files to find the dataset images $\#r$ which contain the same visual phrase. If a ROI from the input image and a ROI from the dataset image

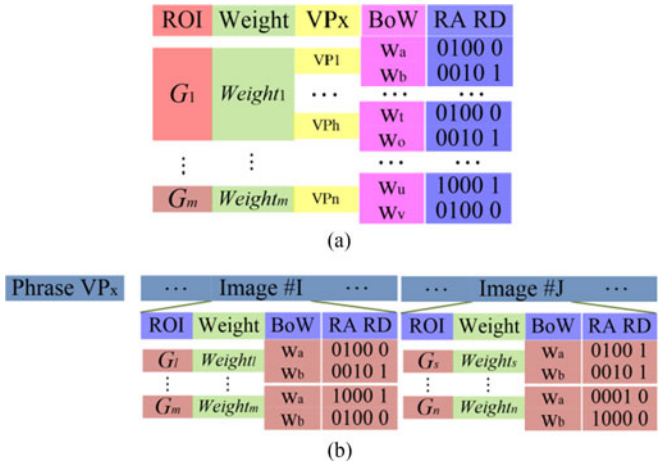


Fig. 3. Illustration of Inverted file structure for the query image and dataset image. G_l is the lth ROI that w_x belongs to. RA and RD are the position descriptors of w_x in the ROI G_l, $Weight_l$ is the weight of the ROI. (a) ROIs and their saliency description for the input image. (b) ROIs and their saliency description for the dataset images.

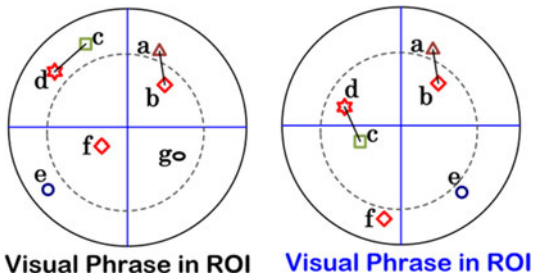


Fig. 4. Visual phrase representation for ROIs from query image and a dataset image.

contain common visual phrase, we call the two ROIs as a matched region pair (denoted by RP). Assuming that the input image and the dataset image have F RPs, denote as RP_f $(f = 1, \ldots, F)$. Based on the RPs, we can measure the similarity of input image and dataset image.

1) ROI Pair Similarity Measurement: For a $RP_f (f = 1, \ldots, F)$, the ROIs from the query image $\#q$ and the dataset image $\#r$ are denoted by G_i^q and G_j^r. Assuming that G_i^q and G_j^r have m common visual phrases $vp_k, (k = 1, 2, \ldots, m)$. The corresponding spatial descriptors for two visual words of vp_k in G_i^q are denoted as SQ_k^1 and SQ_k^2 respectively. And the spatial descriptors for two visual words of vp_k in G_j^r are denoted as SR_k^1 and SR_k^2 respectively. The matching score of a RP is determined based on the geometric consistency of their common visual phrases as follows:

$$MS_f = \sum_{k=1}^{m} \exp\left(-\|SD_k^1 \oplus SR_k^1 + SQ_k^2 \oplus SR_k^2\|\right); \quad f = 1, \ldots, F \tag{6}$$

where $\oplus$ is Logical Exclusive (XOR) operation. If two regions share more common visual phrases with similar spatial distributions, then their matching score will be higher.

Here we give an example, as shown in Fig. 4, where the two ROIs share two common visual phrase, namely (a, b) and (c, d). So, the two ROIs construct a matched region pair. From Fig. 4, we find that the visual phrase (a, b) are in the same locations in the region, while (c, d) are in different locations, their RA and RD descriptors are with large difference, thus the matching score of (a, b) is higher than the visual phrase (c, d).

2) Similarity Measurement: How to measure the similarity of the two images is the key problem of region based image retrieval. There exist one-to-many matched RP_f, because we only determine the ROIs based on whether sharing common visual phase. In our similarity measurement, we fuse the weights to determine the similarity of query image $\#q$ and the dataset image $\#r$ as follows:

$$S_r = \sum_{f=1}^{F} \text{weight}_f^q \times \text{weight}_f^r \times MS_f \tag{7}$$

where weight_f^q and weight_f^r the weights of the RP_f for the query and dataset images. The larger S_r means that the dataset image $\#r$ is more similar with the input image.

Then the images are ranked according to (7) for dataset images. *k*NN based approach is necessary for improving the location estimation performance[4], [35]. The top ranked *k* images are selected. And then we count the number of images for each occurred location. The majority location in the *k* images is assigned for the input image. In this paper, we utilize k = 50 as suggested in [4], [35].

IV. Experimentation

Experiments are done on two datasets: OxBuild, and GOLD [4]. OxBuild is used for preliminary tests. 100 images are selected randomly as the test set, while the rest is served as training set in the offline system. GOLD contains more than 3.3 million images together with their geo-tags. And it covers 60K different cities in the world. 80 travel spots are selected for testing. The test dataset consists of randomly selected 5000 images. Experiments are carried out on a PC with 48G memory and Intel Core(TM)2, Quad CP Q8400 with 2.26 GHz on Matlab.

In order to test the performance of the proposed location estimation approach, comparisons are made with IM2GPS [9], CS [4], SC [8], MSER [36], WSA [24], VWG [35] and the proposed multi-saliency enhancement based image location estimation approach (denoted by SEN).

VWG: visual words mining and spatial constraint based image geographical location estimation approach is exploited. Salient features and spatial information are extracted from the selected useful visual words. Mean-shift clustering is utilized to find visual word group (i.e. ROI in this paper), and then we carry out spatial description for each VWG. The coordinates of BoWs are utilized to carry out geometric consistency checking. Group based spatial coding is conducted. We generate a position descriptor for each visual word. Finally, we carry out image retrieval by region based similarity measurement.

IM2GPS: we utilize the global color and texture feature to estimate the location for the input image. As for IM2GPS, we use the best parameters provided in the paper [9].

CS: This approach consists of two parts: an offline system and an online system. In the offline system, a hierarchical structure is constructed for a large-scale offline social image set with GPS information. Representative images are selected for each GPS location refined cluster, and an inverted file structure is proposed. In the online system, when given an input image, its GPS information can be estimated by hierarchical global clusters selection and local feature refinement in the online system. We utilize hierarchical global feature clustering and local feature refinement under cosine similarity based measurement to find the location for the input image.

SC: spatial coding based approach is carried out for the input image to search similar image search. From the location of the top ranked images, we estimate the location of the query image.

MSER: we utilize the salient regions that extracted by using maximally stable extremal regions, which are viewed as the ROI. And then we also carry out spatial description for the region to find the location of the input image. In this paper, the input of MSER is the same as VWG, i.e. the useful features after selection [35]. The only difference is that we utilize maximally stable extremal region [36] rather than that of our mean-shift based region detection approach to bundle visual words into groups. The other parts are the same as VWG.

WSA: We adopt word spatial arrangement for the ROI that detected by VWG. The only difference is that, in WSA, we adopt word spatial arrangement [24] rather than the position descriptor in ROI. In this paper, the other parts of WSA are also the same as these of VWG. The only difference is that, in WSA, we adopt word spatial arrangement [24] rather than the position descriptor in VWG.

SEN: the proposed approach by saliency enhancement and visual phrase extraction. Compared with VWG, this approach has additional two saliency enhancements: the salient map and the visual phrase expression.

A. Performance Evaluation

For an input image, if the estimated location is exact with its ground-truth (manually labeled location), it is correctly estimated, otherwise falsely estimated. We utilize the average recognition rate (AR) [4] to evaluate image location estimation performance which is given as follows:

$$AR = \frac{1}{G}\sum_{i=1}^{G} RR_i \tag{8}$$

where Z is the number of locations and RR_i is the recognition rate of the i-th location.

$$RR_i = \frac{NC_i}{NI_i} \times 100\%, i \in \{1, 2, \ldots, Z\} \tag{9}$$

where NC_i is the correct estimated image number, NI_i is the test image number.

B. Performance Comparison

The location estimation performances of IM2GPS, SC, CS, MSER, WSA, VWG and SEN are shown in Fig. 5. We find that VWG and SEN outperform the other methods on the two datasets. The AR values of SEN on OxBuild and GOLD are 91.5% and 85.46% respectively. The results of IM2GPS in the two test datasets are 55.5% and 42.16% respectively, which are with lowest performance. This shows that only on the global feature, the location estimation performances are not satisfying. SC achieves 68.5% and 60.428% on the two datasets. Although SC utilizes local feature, it neglects the clues that global feature can provide. While the results of our previous approach CS are 84% and 76.45% respectively. The performance of CS is better than IM2GPS and SC. It shows that both image global and local visual features are beneficial in image location estimation.

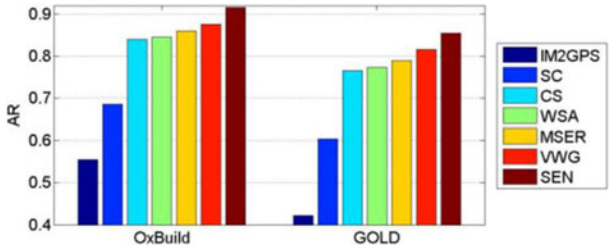


Fig. 5. Performances of IM2GPS, SC, CS, WSA, MSER, VWG, and the proposed SEN-based image location approaches.

TABLE I
AVERAGE COMPUTATIONAL COSTS OF SC, IM2GPS, CS, WSA, MSER, VWG, AND OUR APPROACH SEN

Dataset	SC	IM2GPS	CS	WSA	MSER	VWG	SEN
OxBuild	5 ms	34 ms	0.5 ms	0.40 s	1.54 s	0.38 s	0.58 s
GOLD	47 ms	64 s	0.96 ms	0.64 s	2.01 s	0.62 s	1.16 s

MSER bundles visual words into groups. The AR values of MSER are 86.5% and 78.86%. We divide the useful visual words of an image into VWGs by mean-shift clustering. This shows that our salient region mining approach is effective. In WSA, word spatial arrangement is utilized to encode the distribution of a useful visual word in an image. And the average recognition rates of WSA are 84.5% and 76.65%. Its results are not as good as our spatial coding approach in VWG, which have the performances 87.5% and 81.52% on OxBuild and GOLD respectively. By the comparisons we find that the proposed approach with multi-saliency enhancement is with the best performances, which still achieve 4% and 3.94% over our previous approach VWG on OxBuild and GOLD datasets respectively.

Correspondingly, the computational costs (in second:s or micro-second: ms) of the compared approaches are shown in Table I. The region based approaches such as WSA, MSER, VWG and SEN are all computational intensive than the non-region based approaches. The computational cost of our approach SEN is a bit heavy than WSA and VWG, but still lower than MSER. This is due to the fact that the MSER based salient region detection approach is far more computational heavy than our mean-shift based visual word clustering approach. Compared with VWG, we need to carry out saliency map detection for the input query image and visual phrase generation and searching in SEN. So, its computational cost is corresponding high. In this paper, we do not consider the fast implement for

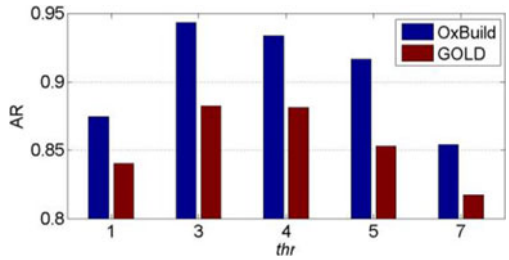


Fig. 6. Impact of thr to location estimation performance.

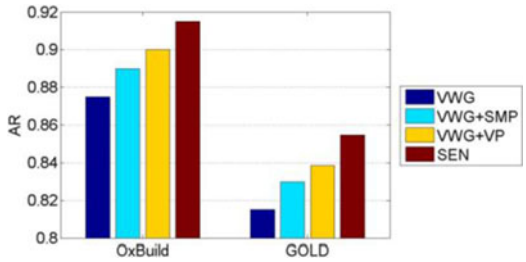


Fig. 7. Contribution of each component in location estimation.

TABLE II
COMPARISON OF AVERAGE COMPUTATIONAL COSTS (S)

Dataset	with indexing	without indexing
OxBuild	0.58	211.53
GOLD	1.16	859.29

our SEN approach. As our approach is based on the mined ROIs, it can be carried out by parallel. This will speed up the computational process in real time image location estimation.

C. Discussion

Hereinafter, we discuss the impacts of some parameters to image location estimation performances. We also show the effect of using image indexing or not.

1) The Impact of thr: In the region based visual phrase extraction, the visual words of the region are considered a phrase only if the distance between the two visual words is less than the given threshold thr. Here, we discuss the impact of thr to location estimation performance. If the thr is too large, then two visual words will not be regarded as a visual phrase. This will affect the accuracy of retrieval. Fig. 6 shows that with the increase of thr, the AR first increases and then into declines. Better performances are got when thr is set to be 3 or 4. This is mainly caused by the fact that when thr is smaller, the two visual words that constructed a visual phrase sometime with similar position descriptors that are robust to distribution variant.

2) The Impact of Indexing: In order to show the efficiency of the fast image indexing approach, we provide a complete comparison under the cases with image indexing and without indexing. The corresponding computational costs (in second:s) are shown in Table II respectively. We find that when building fast indexing structure for dataset images, our algorithm with indexing is efficient than the approach without indexing. The computational cost of utilizing fast indexing structure is less than 0.5% of that of without indexing. Because the without indexing approach requires to carry out similarity measurement by image level feature matching.

3) The Contribution of Each Part: In our proposed saliency enhanced image location estimation approach, we fuse multi-saliency that explored from region of interest (VWG), salient map (in short SMP) of image, and visual phrase (in short VP) to improve image location estimation. Here we give a comprehensive comparison for each component. Fig. 7 shows the corresponding contribution of each component. In Fig. 7, VWG represents the only utilizing the ROI for location estimation, i.e. the visual word group. VWG + VP represents the approach utilizing the visual phrase for location estimation, VWG + SMP corresponds to the approach that salient map is enforced on the ROI, and SEN is actually the combinations of the VWG, visual phrase and salient map, i.e. SEN = VWG + VP + SMP.

In the VWG based approach, we only utilize the spatial descriptors to represent each visual word in a region. This shows that when utilizing the visual phrase the saliency can be improved and the discrimination power of visual phrase is stronger than utilizing individual visual words. By comparison we find that, based on ROI based approach, when utilize the saliency extracted from visual phrase, about 2.5% improvement can be achieved.

When the salient map information is utilized (i.e. VWG + SMP), we find that the performance is improved by about 1.5% over ROI. This shows that the visual saliency that detected from the query image and dataset images has positive contribution to reduce unimportant regions to find accurate location. By making full utilize of the three components, better performances can be achieved.

V. CONCLUSION

In this paper, we show that the enhanced saliency information from the region of interest is helpful for improving the geo-location inference for images. The saliency explored from the salient map is also helpful for rejecting some irrelevant region, thus it gives some positive contributions in location estimation. Furthermore, the visual phrase still makes the saliency more positive. Our approach with indexing structure is with about 500 times faster than that without indexing structure. However, it is still a bit computational intensive than the-state-of-the-art approaches. Considering that the ROI based feature representation and similarity measurement can be processed parallel, we can speed up the process by parallel computing. The fast indexing structure is important for reducing the computational costs. While with popularity of deep learning based approach in visual search and recognition. In our future, we will explore saliency from the deep feature learning rather than the handicraft features that we utilized in this paper.

REFERENCES

[1] S. Zhang, Q. Tian, G. Hua, Q. Huang, and S. Li, "Descriptive visual words and visual phrases for image," in *Proc. 17th ACM Int. Conf. Multimedia*, 2009, pp. 75–84.

[2] Y. Li, D. Crandall, and D. Huttenlocher, "Landmark classification in large-scale image collections," in *Proc. 12th Int. Conf. Comput. Vis.*, 2009, pp. 1957–1964.

[3] M. Donoser and D. Schmalstieg, "Discriminative feature to point matching in image based localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2014, pp. 516–523.

[4] J. Li, X. Qian, Y. Tang, L. Yang, and T. Mei, "GPS estimation for places of interest from social users' uploaded photos," *IEEE Trans. Multimedia*, vol. 15, no. 8, pp. 2058–2071, Dec. 2013.

[5] C. Hauff and G. Houben, "Placing images on the world map: A microblog-based enrichment approach," in *Proc. 35th Int. ACM SIGIR Conf. Res. Dev. Inf. Retrieval*, 2012, pp. 691–700.

[6] H. Liu, T. Mei, J. Luo, H. Li, and S. Li, "Finding perfect rendezvous on the go: Accurate mobile visual localization and its applications to routing," in *Proc. 20th ACM Int. Conf. Multimedia*, 2012, pp. 9–18.

[7] E. Gavves, C. Snoek, and A. Smeulders, "Visual synonyms for landmark image retrieval," *Comput. Vis. Image Understand.*, vol. 116, no. 2, pp. 238–249, 2011.

[8] W. Zhou, Y. Lu, H. Li, Y. Song, and Q. Tian, "Spatial coding for large scale partial-duplicate web image search," in *Proc. 18th ACM Int. Conf. Multimedia*, 2010, pp. 511–520.

[9] J. Hays and A. Efros, "IM2GPS: Estimating geographic information from a single image," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2008, pp. 1–8.

[10] O. Chum, J. Philbin, J. Sivic, M. Isard, and A. Zisserman, "Total recall: Automatic query expansion with a generative feature model for object retrieval," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2007, pp. 1–8.

[11] J.Wang, J.Yang, and K.Yu, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recog.*, Jun. 2006, vol. 2, pp. 2169–2178.

[12] X. Qian, H. Wang, G. Liu, and X. Hou, "HWVP: Hierarchical wavelet packet texture descriptors and their applications in scene categorization and semantic concept retrieval," *Multimedia Tools Appl.*, vol. 69, no. 3, pp. 897–920, 2014.

[13] G. Baatz, K. Köser, D. Chen, R. Grzeszczuk, and M. Pollefeys, "Leveraging 3D city models for rotation invariant place-of-interest recognition," *Int. J. Comput. Vis.*, vol. 96, no. 3, pp. 315–334, 2012.

[14] Y. Xue and X. Qian, "Visual summarization of landmarks via viewpoint modeling," in *Proc. IEEE Int. Conf. Image Process.*, Sep.-Oct. 2012, pp. 2873–2876.

[15] J. Li, X. Qian, Y. Tang, L. Yang, and C. Liu, "GPS estimation from users' photos," in *Proc. 19th Int. Conf. Multimedia Model.*, 2013, pp. 118–129.

[16] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Object retrieval with large vocabularies and fast spatial matching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2007, pp. 1–8.

[17] C. Wu, F. Fraundorfer, J. Frahm, and M. Pollefeys, "3D model search and pose estimation from single images using VIP features," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recog. Workshops*, Jun. 2008, pp. 1–8.

[18] Y. Zheng *et al.*, "Tour the world: Building a web-scale landmark recognition engine," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2009, pp. 1085–1092.

[19] 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.

[20] M. Park, J. Luo, R. Collins, and Y. Liu, "Beyond GPS: Determining the camera viewing direction of a geo-tagged image," in *Proc. 18th ACM Int. Conf. Multimedia*, 2010, pp. 631–634.

[21] D. Crandall, L. Backstrom, D. Huttenlocher, and J. Kleinberg, "Mapping the world's photos," in *Proc. Int. Conf. World Wide Web*, 2009, pp. 761–770.

[22] E. Kalogerakis, O. Vesselova, J. Hays, A. Efros, and A. Hertzmann, "Image sequence geolocation with human travel priors," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, Sep.-Oct. 2009, 253–260.

[23] T. Quack, B. Leibe, and L. Gool, "World-scale mining of objects and events from community photo collections," in *Proc. Int. Conf. Content-Based Image Video Retrieval*, 2008, pp. 47–56.

[24] O. Penatti, F. Silva, and E. Valle, "Visual word spatial arrangement for image retrieval and classification," *Pattern Recog.*, vol. 47, no. 2, pp. 705–720, 2014.

[25] Y. Zhao, X. Qian, and T. Mu, "Image taken place estimation via geometric constrained spatial layer matching," in *Proc. 21st Int. Conf. MultiMedia Model.*, 2015, pp. 436–446.

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

[27] J. Han, M. Xu, X. Li, L. Guo, and T. Liu, "Interactive object-based image retrieval and annotation on iPad," *Multimedia Tools Appl.*, vol. 72, no. 3, pp. 2275–2297, 2014.

[28] S. Jiang, X. Qian, J. Shen, Y. Fu, and T. Mei, "Author topic model-based collaborative filtering for personalized POI recommendations," *IEEE Trans. Multimedia*, vol. 17, no. 6, pp. 907–918, Jun. 2015.

[29] O. Laere, S. Schockaert, and B. Dhoedt, "Ghent university at the 2010 placing task," in *Proc. MediaEval Workshop*, 2010, pp. 1–2.

[30] P. Kelm, S. Schmiedeke, and T. Sikora, "How spatial segmentation improves the multimodal geo-tagging," in *Proc. MediaEval*, 2012, pp. 1–2.

[31] S. Zhang *et al.*, "Building contextual visual vocabulary for large-scale image applications," in *Proc. 18th ACM Int. Conf. Multimedia*, 2010, pp. 501–510.

[32] H. Li, Y. Wang, T. Mei, J. Wang, and S. Li, "Interactive multimodal visual search on mobile device," *IEEE Trans. Multimedia*, vol. 15, no. 3, pp. 594–607, Apr. 2013.

[33] J. Sang *et al.*, "Interaction design for mobile visual search," *IEEE Trans. Multimedia*, vol. 15, no. 7, pp. 1665–1676, Nov. 2013.

[34] D. Nistér and H. Stewénius, "Scalable recognition with a vocabulary tree," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2006, vol. 2, pp. 2161–2168.

[35] X. Qian, Y. Zhao, and J. Han, "Image location estimation by salient region matching," *IEEE Trans. Image Process.*, vol. 24, no. 6, pp. 4348–4358, Nov. 2015.

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

[37] J. Chen, B. Feng, L. Zhu, P. Ding, and B. Xu, "Effective near-duplicate image retrieval with image-specific visual phrase selection," in *Proc. IEEE Int. Conf. Image Process.*, Sep.-Oct. 2010, pp. 1909–1912.

[38] J. Li *et al.*, "Improved image GPS location estimation by mining salient features," *Signal Process., Image Commun.*, vol. 38, pp. 141–150, 2015.

[39] R. Ji, L. Duan, J. Chen, and W. Gao, "Towards compact topical descriptor," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2012, pp. 2925–2932.

[40] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Lost in quantization: Improving particular object retrieval in large scale image databases," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2008, pp. 1–8.

[41] W. Tang, R. Cai, Z. Li, and L. Zhang, "Contextual synonym dictionary for visual object retrieval," in *Proc. 19th ACM Int. Conf. Multimedia*, 2011, pp. 503–512.

[42] S. Jiang, X. Qian, T. Mei, and Y. Fu, "Personalized travel sequence recommendation on multi-source big social media," *IEEE Trans. Big Data*, vol. 1, no. 2, pp. 43–56, Mar. 2016.

[43] D. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

[44] Y. Yang *et al.*, "A multimedia retrieval framework based on semi-supervised ranking and relevance feedback," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 4, pp. 723–742, Apr. 2012.

[45] L. Shao, L. Liu, and X. Li, "Feature learning for image classification via multi-objective genetic programming," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 7, pp. 1359–1371, Jul. 2014.

[46] Y. Yang, Z. Ma, A. Hauptmann, and N. Sebe, "Feautre selection for multimedia analysis by sharing information among multiple tasks," *IEEE Trans. Multimedia*, vol. 15, no. 3, pp. 661–669, Apr. 2013.

[47] C. Yang, J. Shen, J. Peng, and J. Fan, "Image collection summarization via dictionary learning for sparse representation," *Pattern Recog.*, vol. 46, no. 3, pp. 948–961, 2013.

[48] W. Min, B. Bao, and C. Xu, "Multimodal spatio-temporal theme modeling landmark analysis," *IEEE Multimedia Mag.*, vol. 21, no. 3, pp. 20–29, Jul./Sep. 2014.

[49] J. Huang, H. Liu, J. Shen, and S. Yan, "Towards efficient sparse coding for scalable image annotation," in *Proc. 21st ACM Int. Conf. Multimedia*, 2013, pp. 947–956.

[50] H. Bay, T. Tuytelaars, and L.Gool, "Surf: Speeded up robust features," in *Proc. 9th Eur. Conf. Comput. Vis.*, 2006, pp. 404–417.

[51] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 11, pp. 1254–1259, Nov. 1998.

[52] M. Cheng, N. Mitra, X. Huang, P. Torr, and S. Hu, "Global contrast based salient region detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 3, pp. 569–582, Mar. 2015.

[53] H. Li and K. N. Ngan, "A co-saliency model of image pairs," *IEEE Trans. Image Process.*, vol. 20, no. 12, pp. 3365–3375, Dec. 2011.

[54] X. Cao, Z. Tao, B. Zhang, H. Fu, and W. Feng, "Self-adaptively weighted co-saliency detection via rank constraint," *IEEE Trans. Image Process.*, vol. 23, no. 9, pp. 4175–4186, Sep. 2014.
[55] D. Zhang, J. Han, C. Li, and J. Wang, "Co-saliency detection via looking deep and wide," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2015, pp. 2994–3002.
[56] H. Fu, X. Cao, and Z. Tu, "Cluster-based co-saliency detection," *IEEE Trans. Image Process.*, vol. 22, no. 10, pp. 3766–3778, Oct. 2013.
[57] X. Cao, Y. Cheng, Z. Tao, and H. Fu, "Co-saliency detection via base reconstruction," in *Proc. 22nd ACM Int. Conf. Multimedia*, 2014, pp. 997–1000.
[58] X. Yang, X. Qian, and T. Mei, "Learning salient visual word for scalable mobile image retrieval," *Pattern Recog.*, vol. 48, no. 10, pp. 3093–3101, 2015.
[59] X. Qian, Y. Xue, Y. Tang, X. Hou, and T. Mei, "Landmark summarization with diverse viewpoints," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 25, no. 11, pp. 1857–1869, Nov. 2015.
[60] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1106–1114.
[61] X. Qian, X. Liu, X. Ma, D. Lu, and C. Xu, "What Is happening in the video? Annotate video by sentence," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 26, no. 9, pp. 1746–1757, Sep. 2016.
[62] R. Xia, Y. Pan, H. Lai, C. Liu, and S. Yan, "Supervised hashing via image representation learning," in *Proc. 28th AAAI Conf. Artif. Intell.*, 2014, pp. 2156–2162.
[63] R. Zhang, L. Lin, R. Zhang, W. Zuo, and L. Zhang, "Bit-scalable deep hashing with regularized similarity learning for image retrieval and person re-identification," *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 4766–4779, Dec. 2015
[64] K. Lin, H. Yang, J. Hsiao, and C. Chen, "Deep learning of binary hash codes for fast image retrieval," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog. Workshops*, Jun. 2015, pp. 27–35.
[65] F. Zhao, Y. Huang, L. Wang, and T. Tan, "Deep semantic ranking based hashing for multi-label image retrieval," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2015, pp. 1556–1564.
[66] D. Lu, X. Liu, and X. Qian, "Tag based image search by social re-ranking," *IEEE Trans. Multimedia*, vol. 18, no. 8, pp. 1628–1639, Aug. 2016.
[67] G. Zhao, X. Qian, and T. Mei, "Service rating prediction by exploring social mobile users' geographic locations," *IEEE Trans. Big Data*, to be published, doi: 10.1109/TBDATA.2016.2552541.
[68] X. Qian, X. Tan, Y. Zhang, R. Hong, and M. Wang, "Enhancing sketch-based image retrieval by re-ranking and relevance feedback," *IEEE Trans. Image Process.*, vol. 25, no. 1, pp. 195–208, Jan. 2016.
[69] X. Lu, Y. Yuan, X. Zheng, "Jointly dictionary learning for change detection in multispectral imagery," *IEEE Trans. Cybern.*, to be published, doi: 10.1109/TCYB.2016.2531179.
[70] X. Lu, Y. Yuan, and P. Yan, "Image super-resolution via double sparsity regularized manifold learning," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 23, no. 12, pp. 2022–2033, Dec. 2013.
[71] X. Lu, Y. Yuan, and P. Yan, "Alternatively constrained dictionary learning for image super-resolution," *IEEE Trans. Cybern.*, vol. 44, no. 3, pp. 366–377, Mar. 2014.
[72] X. Lu and X. Li, "Multi-resolution imaging," *IEEE Trans. Cybern.*, vol. 44, no. 1, pp. 149–160, Jan. 2014.
[73] X. Lu X. Li, and L. Mou, "Semi-supervised multi-task learning for scene recognition," *IEEE Trans. Cybern.*, vol. 45, no. 9, pp. 1967–1976, Sep. 2015.
[74] X. Lu, Z. Wang, and X. Lu , "Surveillance video synopsis via scaling down objects," *IEEE Trans. Image Process.*, vol. 25, no. 2, pp. 740–755, Feb. 2016.
[75] X. Lu, X. Zheng, and X. Li, "Latent semantic minimal hashing for image retrieval," *IEEE Trans. Image Process.*, vol. 26, no. 1, pp. 355–368, Jan. 2017.
[76] X. Lu, L. Mou, and X. Lu, "Scene recognition by manifold regularized deep learning architecture," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 10, pp. 2222–2233, Oct. 2015.
[77] D. Zhang, J. Han, C. Li, J. Wang, and X. Li, "Detection of co-salient objects by looking deep and wide," *Int. J. Comput. Vis.*, vol. 120, no. 2, pp. 215–232, 2016.
[78] Y. Zhang, X. Qian, X. Tan, J. Han, and Y. Tang, "Sketch-based image retrieval by salient contour reinforcement," *IEEE Trans. Multimedia*, vol. 18, no. 8, pp. 1604–1615, Aug. 2016.
[79] D. Zhang, J. Han, J. Han, and L. Shao, "Cosaliency detection based on intrasaliency prior transfer and deep intersaliency mining," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 6, pp. 1163–1176, Jun. 2016.

Xueming Qian (M'09) received the B.S. and M.S. degrees from the Xi'an University of Technology, Xi'an, China, in 1999 and 2004, respectively, and the Ph.D. degree from the School of Electronics and Information Engineering, Xi'an Jiaotong University, Xi'an, China, in 2008.

He was a Visiting Scholar with Microsoft Research Asia, Beijing, China, from 2010 to 2011. He was previously an Assistant Professor with Xi'an Jiaotong University, where he was an Associate Professor from 2011 to 2014, and is currently a Full Professor. He is also the Director of the Smiles Laboratory, Xi'an Jiaotong University. His research is supported by the National Natural Science Foundation of China, Microsoft Research, and the Ministry of Science and Technology. His research interests include social media big data mining and search.

Prof. Qian was the recipient of the Microsoft Fellowship in 2006. He was the recipient of the Outstanding Doctoral Dissertations Award of Xi'an Jiaotong University and Shaanxi Province in 2010 and 2011, respectively.

Huan Wang received the B.S. degree from Xi'an University of Technology, Xi'an, China, in 2004, and the M.S. degree in 2010 from Xi'an Jiaotong University, Xi'an, China, where she is currently working toward the Ph.D. degree.

Her research interests include video/image coding, communication, and transmission.

Yisi Zhao is currently working toward the M.S. degree at the Smiles Laboratory, Xi'an Jiaotong University, Xi'an, China.

Her research interests include large-scale image retrieval and image content understanding.

Xingsong Hou received the Ph.D. degree from the School of Electronics and Information Engineering, Xi'an Jiaotong University, Xi'an, China, in 2005.

From 1995 to 1997, he was an Engineer with the Xi'an Electronic Engineering Institute, Xi'an, China, in the field of radar signal processing. He is currently a Professor with the School of Electronics and Information Engineering, Xi'an Jiaotong University. His research interests include video/image coding, wavelet analysis, sparse representation, sparse representation and compressive sensing, and radar signal processing.

Richang Hong (M'14) received the Ph.D. degree from the University of Science and Technology of China, Hefei, China, in 2008.

He was a Research Fellow with the School of Computing, National University of Singapore, Singapore, from 2008 to 2010. He is currently a Professor with the Hefei University of Technology, Hefei, China. He has coauthored more than 60 publications in the areas of his research interests, which include multimedia question answering, video content analysis, and pattern recognition.

Prof. Hong is a Member of the Association for Computing Machinery. He was the recipient of the Best Paper Award in ACM Multimedia 2010.

Meng Wang (M'09) received the B.E. and Ph.D. degrees from the Department of Electronic Engineering and Information Science, University of Science and Technology of China, Hefei, China.

He is currently a Professor with the Hefei University of Technology, Hefei, China. His current research interests include multimedia content analysis, search, mining, recommendation, and large-scale computing.

Yuan Yan Tang (F'04) received the B.E. degree in electrical and computer engineering from Chongqing University, Chongqing, China, in 1966, the M.Eng. degree in electrical engineering from the Beijing Institute of Post and Telecommunications, Beijing, China, in 1981, and the Ph.D. degree in computer science from Concordia University, Montreal, QC, Canada, in 1990.

He is the Chair Professor with the Faculty of Science and Technology, University of Macau, Macau, China, and a Professor, an Adjunct Professor, or Honorary Professor at several institutes, including Chongqing University, Chongqing, China; Concordia University, Portland, OR, USA; and the Hong Kong Baptist University, Hong Kong. His research interests include wavelet theory and applications, pattern recognition, image processing, document processing, artificial intelligence, and Chinese computing.