Signal Processing: *Image Communication* 38 (2015) 141–150

Contents lists available at ScienceDirect

Signal Processing: *Image Communication*

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

Improved image GPS location estimation by mining salient features ☆

Jing Li [a,b], Xueming Qian [a,*], Ke Lan [a,c], Peilun Qi [a], Arunabh Sharma [b]

[a] *28 Xianning West Road, Xi'an Jiaotong University, China*
[b] *475 Northwestern Ave, Purdue University, USA*
[c] *2201 West End Ave, Vanderbilt University, USA*

ARTICLE INFO

Available online 26 July 2015

Keywords:
GPS estimation
Social media
Salient region
Salient feature
Image groups
BoW
Geo-tag

ABSTRACT

Nowadays, people tend to share their personal photos, taken while they are traveling, to the social media sharing websites, such as Flickr. There is also convenient access to the large scale image dataset, usually attached with metadata such as GPS location, tags and description so on. With the help of images taken in places of interest in conjunction with the broad multimedia information realm, the task of automatic image GPS location estimation became possible. However, automatic image GPS location estimation is still a nontrivial task even in today's world with explosive quantity of images available on the website. In general, images taken from identical locations share some features, such as some salient features, even when the images are taken from different viewpoints. These salient features play an important role in the image location estimation. Thus, in this paper, we propose a salient image feature mining based image GPS location estimation method. We 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 relation with the neighbor image groups. Experiments on different datasets demonstrate the effectiveness of the proposed method.

1. Introduction

GPS information is useful in image recognition, automatic travel guidance and recommendation [1–8]. The task of image GPS location estimation becomes feasible with the aid of an explosive growth of geo-tagged images on social media sharing networks. Large scale geo-tagged image datasets are available on websites such as Flickr and Picasa. Nowadays, how to recognize geo-referenced images automatically draws more and more attention [1–5,9]. GPS information of social images has been widely used in content browsing [1–3,10–12], image annotation [13–15], image search [16–18], and localization [9,19]. For example, full view of a landmark can be realized by building 3-D models by collecting large-scale geo-tagged photos [20–22]. Geo-location based image retrieval is carried out by generating 3-D models and translating the query image into a 3-D pattern [19–21]. However, theoretically and practically, 3D reconstruction is not a trivial task, it depends much on the image quality and the information of the camera. In fact, images uploaded by users to the social websites' even with really high quality

☆ This work is supported in part by the Program 973 No. 2012CB316400, by NSFC Nos. 60903121, 61173109, 61202180, and 61332018, Microsoft Research Asia, and Fundamental Research Funds for the Central Universities.

* Corresponding author.
E-mail addresses: li1463@purdue.edu (J. Li), qianxm@mail.xjtu.edu.cn (X. Qian).

http://dx.doi.org/10.1016/j.image.2015.07.007
0923-5965/

usually lack information needed in 3D reconstruction, which makes it difficult to use the method of 3D reconstruction in our goals of solving the problem of image GPS location estimation based on the large scale social image sets. Qian et al. have shown that using the GPS information of user uploaded photos is helpful for improving users vocabulary tagging performances [6]. Moreover, the Placing Task makes use of attached metadata, such as tags and user descriptions to estimate the GPS location of an image or a video frame [3–5].

A lot of research effort has been devoted to image GPS estimation [1,2,10,9,19]. Generally, GPS location estimation can be achieved by means of image matching [1], image retrieval [18,9], and image classification [23]. The main process can be as follows: first, find images similar to the input image, then assign the GPS location of the visually similar images to that of the input image. From this point of view, existing example-based image retrieval approaches can be utilized in GPS location estimation for an input image [2,17,18]. Kelm et al. provide a method of video localization by taking advantage of textual and visual modalities. They tackle the geo-referencing problem with a hierarchical classification approach. The world map is divided into segments of different sizes and each segment is considered as a class for the probabilistic model. For the probabilistic model, textual and visual approaches are provided. They adopt a centroid-based candidate fusion to solve the problem of data sparsity and enhance the distributions of single candidates in a multi-modal manner. For visual information, global features such as color and texture are utilized [24]. Refs. [43,44] focus on landmarks, providing both content based landmark classification and landmark search algorithms. In these approaches, each image is represented by a set of global/local low-level features. Due to the gap between image and its low level visual feature descriptors, image retrieval is still very challenging [25]. Much attentions have also been paid to effective image feature representations [26–30]. Global features usually neglect the local saliency in the image, even though a lot of scalable or rotation invariant features are utilized [28,29]. Therefore much research focus on local feature matching based image retrieval [26,27]. However, feature matching based approaches is both low accuracy and has a heavy computational cost for a large scale dataset. To solve the problem, bag-of-word (BoW) model based approaches are proposed [31–38]. In these approaches, fast image retrieval is achieved by giving each BoW a weight and then computing scores for each image in the dataset. Thus, the well-known TF-IDF [12,13] (term frequency inverse document frequency) which often utilizes text based retrieval can be adopted to carry out image retrieval. TF-IDF is a numerical statistic which reflects how important a word is to a document in a collection or a corpus [12–14]. It can be viewed as a coarse level salient feature representation approach [39]. It is often used as a weighting factor in information retrieval and text mining. With the popularity of BoW model in image processing [14], TF-IDF has been widely utilized in image retrieval and classification [13,31,32,9,18]. Zhang et al. proposed a spatial coding based image retrieval approach by building the contextual BoW [18]. By using inverted construction of BoW and the spatial constraints, the computational cost is low and the performance is satisfactory. However, TF-IDF has the disadvantage of limited capability in class identification since the computation of the weight neglects the class information. In [34], Zhao et al. proposed to utilize the spatial layer matching to improve image location estimation performance and they also further explored the spatial constraint to carry out location estimation [37].

To speed up the estimation to meet the real-time applications on mobile ends, hierarchical global feature clustering and local feature verification and fast invert file indexing structure approaches were proposed [2,17,33,34,37].

Knopp et al. [40] developed a method for automatically detecting such confusing objects and demonstrated that removing them from the database can significantly improve the place recognition performance. It assumes that an image of a particular place does not match well to other images at far away locations. For all the images in the offline dataset, the confusing scores are computed for the local features in the images. Then in the retrieval step, confusing scores are taken into consideration. Instead of computing confusing scores for the confusing features, our main goal is to mine salient features that are critical to the GPS location of the input image.

Han et al. [41] developed a probabilistic computational algorithm by integrating objectness likelihood with appearance rarity to detect object-oriented visual saliency. Compactness, continuity, and center bias are utilized as the measure for the likelihood of the object. They also propose a framework for saliency detection by modeling the background and then separating salient objects from the background [42]. They formulate the problem of salient object detection as background subtraction problems. Our goal is to extract salient features shared by visually similar images of the same places, and utilize them to improve image location estimation performances. Han et al. [42] propose a framework for saliency detection by first modeling the background and then separating salient objects from the background [36].

The task of content based image location estimation is to select images with similar appearance and using their tagged locations to estimate the GPS location of the input image. To our knowledge, most of the existing image location estimation approaches generally treat the input image independently [1–5,9,34,37] but not considering the relationship with their visually similar images in social media communities (i.e. at cloud/server end). Query expansion is utilized to fuse more relevant information from the iterative search result [15]. Moreover, fusing the visual and textual information of the query can also be helpful for the retrieval more relevant results [25]. However, both query expansion and visuo-textual joint image retrieval use external information to improve image retrieval performances. The intrinsic characters of the input image are not deeply explored. The role of salient parts of input image and their relationships with their nearest neighbors is overlooked. Actually, images taken at same location should share some common content. The salient parts of places-of-interest usually appear in most social users' shared photos. Moreover, for the input image, different parts

have different contributions to the recognition of the place. For example, the background (such as green grass, which may appear frequently in many images) is less contributive, while the salient parts (e.g. the spire of a tower) are more contributive. Therefore in our method, the input image and its K nearest image groups are fused to mine salient features to improve location estimation performances.

We propose a salient feature mining based image location estimation method by exploring the salient part among the neighbor groups and then picking salient features to retrieval images taken at an identical location. The proposed approach consists of the following three steps. First, neighbor image groups selection. We select K-nearest image groups for the input image by utilizing the method of GPS estimation in [9]. Second, we mine the salient features from neighbor image groups. Based on the relation between the input image and K-nearest image groups, the local features in the input image are sorted and the salient features are selected. The details of salient feature definition and selection are described later in the online system overview. Finally, the selected salient features are utilized as queries to the selected image groups from the whole dataset and K-NN is used to determine the final GPS location estimation result. The main contributions of our work are as follows:

- Information provided by the K nearest image groups is jointly utilized in the salient feature selection.
- We utilize the mechanism of feedback to improve GPS estimation performance. With the aid of the selected K nearest image groups from the first iteration, the contributions of each of the salient parts can be measured quantitatively.
- We fuse saliency of each BoW in determining the location. The salient components of an image will be assigned higher weights.

The rest of the paper is organized as follows: Section 2 introduces the system overview of the proposed image GPS estimation approach. The offline and online systems of this paper are given in Sections 3 and 4 respectively. Experiments and discussions are shown in Section 5. Conclusions are drawn in Section 6.

2. System overview

We propose an improved GPS estimation algorithm that uses neighboring image groups of the input image over our preliminary approaches [9]. Compared to our previous algorithm, there are two main differences. The first is to utilize the final selected image groups as an internal result instead of using them to estimate GPS location. And the second is that salient features are mined based on the clues provided by the internal result. The block diagram of our approach is shown in Fig. 1. It consists of online and offline systems.

The offline system is identical to that of [9] as shown in Fig. 1. It aims to index the geo-tagged image dataset. It consists of the following six parts: (1) dataset preprocessing to remove noisy images, (2) feature representation by extracting global and local features, (3) clustering the images into R categories utilizing global features, (4) obtaining GPS location refined centroids (i.e. each centroid corresponds to an identical GPS location) for each first layer cluster, (5) selecting representative images for each refined centroid, (6) building inverted files for the representative images of each refined centroid based on the BoW of the SIFT descriptor. The detailed steps of the offline system are presented in Section 3.

The online system takes advantage of the internal result of our previous GPS estimation [9], as shown in Fig. 1(a). The online system estimates the GPS location of the input image. The online system of the proposed approach as shown in Fig. 1(c) carries out the following seven steps after global and local feature extraction: (1) first layer cluster selection, (2) second layer centroid selection, (3) local feature refinement, (4) K nearest neighbor image groups selection, (5) contributions of local salient parts determination, (6) inverted file referring and similarity measurement, and (7) GPS location estimation and verification. In the first step, the input image is assigned to one or more of the image clusters. Images in the same cluster are visually similar. In the second step, the input image is classified into a number of GPS location refined centroids. In the third step, the local feature refinement improves the accuracy of GPS location estimation. Also, as the local refinement is confined to a much smaller scale compared to the whole dataset, it has a low computation cost. In the fourth step, we determine the K nearest image groups for the input image. In the fifth step, we calculate the contributions of the local salient parts. In the sixth step, we carry out similarity measurement between the input image and all the geo-tagged images in the offline dataset. Finally GPS of the input image is assigned based on the GPS information of the images in its K-nearest neighbors. The detailed steps of the online system are presented in Section 4. The novel part of this paper is the following steps: neighbor image groups selection, salient feature mining and inverted file structure referring and scoring in the online system. The details are discussed in Section 4.

3. The offline system

The offline system of this paper is identical to [9], as shown in Fig. 1(b). We only give a brief illustration for each part of the offline system hereinafter. For more details, turn to [9].

3.1. Preprocessing for the dataset

The aim of this step is to remove the noisy images (too bright or too dark) from the dataset by checking their average luminance and texture energy. These kinds of images have little to contribute to online image GPS location estimation. If the image has a high enough or low enough average energy, or has very low texture energy, then it is viewed as noise and removed from the dataset.

3.2. Feature representation

In this paper, three kinds of low-level features are utilized [9]. They are the 45D color moment (CM), 170D hierarchical wavelet packet descriptor (HWVP), SIFT [27]

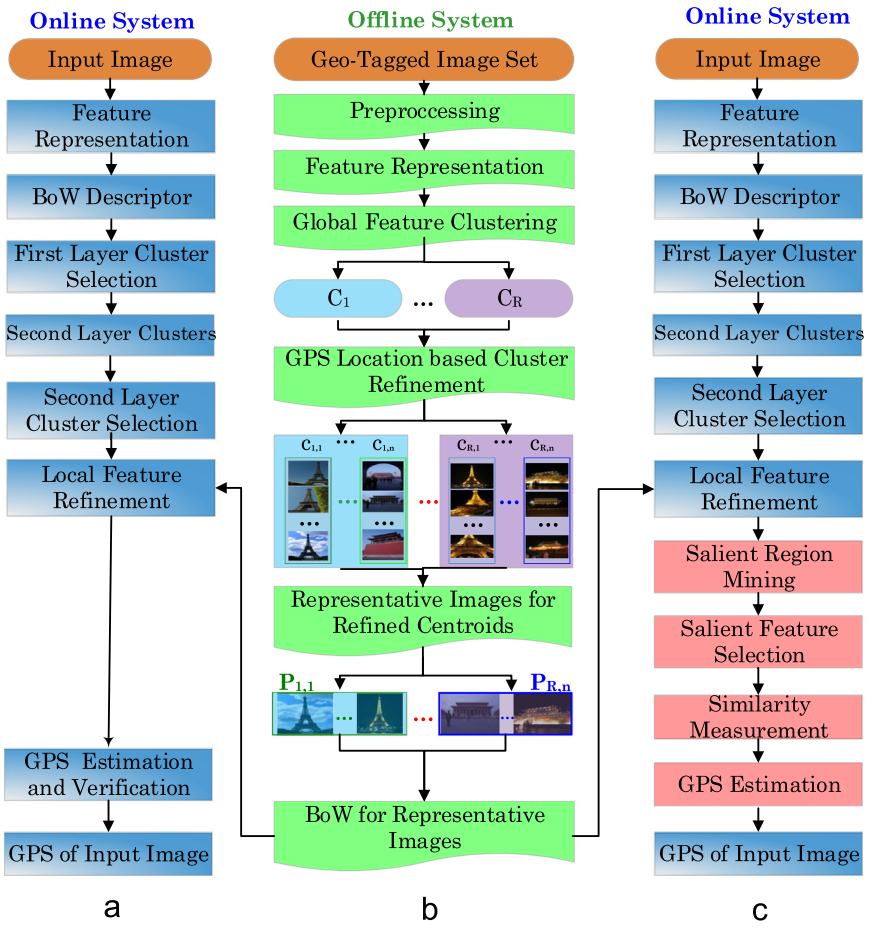

Fig. 1. The flowchart of the proposed salient region mining and salient feature selection based location estimation for an image. (a) Online system of previous model, (b) offline system, (c) online system of our algorithm in this paper.

and its representation in bag-of-visual-word. We quantize each SIFT feature point into a BoW by hierarchical quantization. Each SIFT point is quantized into one of the Q centroids. In this paper, the value of Q is 68,420.

The global features CM and HWVP are utilized in training the hierarchical global feature clusters in the offline system and carry out hierarchical classification for the input image in the online system.

In our preliminary paper, the SIFT-based local feature matching is used for determining: (1) whether or not the input image was taken from an offline GPS location, and (2) in which place the input image was taken. In this paper, SIFT feature matching is utilized in both the offline system for representative image selection for each GPS location refined centroid and the online system to determine the matched representative images.

3.3. Global feature clustering

In this paper, we use K-means to cluster the global features. The number of first layer clusters R in k-means is set to $R=32$ by considering both the image GPS estimation performances and computational costs.

The global feature clustering is carried out on the combined 215-D low-level feature including 45-D color moment and 170-D hierarchical wavelet packet. The global features of all the images in the offline dataset are grouped into R centroids using K-means. After the global feature clustering, we get R centroids $C_1,\ldots,C_R$. Each centroid $C_i(i,=1,\ldots,R)$ is featured by a 215-D global feature vector LC_i.

3.4. GPS location based cluster refinement

After obtaining the set of centroids $\{C_1,\ldots,C_R\}$, we then partition the set of geo-tagged images into these R clusters. We then create sub-clusters based on the GPS locations within each of these clusters, yielding a further partitioning of the images into clusters $c_{ij}(i=1,\ldots,R;j=1,\ldots,N)$.

3.5. Representative images selection for the GPS location refined centroids

The advantage of hierarchical global feature clustering is the low computational cost. In the hierarchical global feature clustering stage, we group images into coarse clusters C_i and refine them into GPS locations refined centroid c_{ij} .

Ideally, the images in the same GPS refined centroid have similar visual content, but actually there are some outliers with incorrect GPS information. Selecting representative images for each GPS location has the following merits: reduced computational cost, and noise suppression to improve GPS estimation performances.

3.6. BoW for representative image groups

For the offline geo-tagged dataset, each SIFT point is quantized into one of the Q centroids. We build an inverted file structure for all the representative images in the offline dataset. The inverted file is a hierarchical structure as detailed in [9]. For the BoW_x, $(x \in 1,\ldots,Q)$, the first layer cluster C_i, the second cluster $c_{i,j}$ and the $image_L$ inside the cluster $c_{i,j}$ are all recorded. In addition, the frequency of the BoW in all the image datasets (denoted as $Frequen_x$) and that in $image_L$ (denoted as $Freq_L(x)$) are also recorded. The number of BoW in $image_L$ ($Number_L$) and the number of image clusters ($Number_C$) in which the BoW occurs are recorded as well and used in the online system.

4. The online system

The online system of the proposed GPS location estimation approach of an input image is shown in Fig. 1(c). First, we extract CM and HWVP, SIFT, and quantize each SIFT point into a BoW for the input image. Then, we carry out hierarchy layer cluster selection and local feature refinement, which are identical to the previous version [9]. The novel part of this paper is the subsequent steps: neighbor image groups selection, salient feature mining and inverted file structure referring and scoring. The last step is image GPS location estimate using image ranking plus K-NN.

The significant difference from our previous work lies in the feature points selection and the computation of weight. Instead of using all the features inside the input image, we develop an algorithm to mine only the features that are more salient for describing the image. The way we measure the saliency of features is by considering the relation between the features in the input image and the selected representative image groups. Also, we modify the traditional TF-IDF weight computation, so that the computed weight of the feature is more representative for the saliency of the feature.

4.1. Hierarchical layer cluster selection

This part consists of the first layer cluster selection and second layer cluster selection. Let L_{input} denote the 215-D global features (consisting of 45-D CM and 170-D HWVP) of the input image. The distance D_i between the query image and the feature vector LC_i of the ith center C_i is computed. The top ranked $M(M \le R)$ centroids in the first layer are selected. Let set $S=\{S_1,\ldots,S_M\}$ denote the selected M candidates, where $S_k \in \{C_1,\ldots,C_R\}$ is one of the selected candidates $(k \in \{1,\ldots,M\})$. In this paper we use $M=10$, in accordance with our previous work. After selecting the first layer cluster candidates set $S=\{S_1,\ldots,S_M\}$, the input image can be further refined into the second layer GPS location refined centroids. Each $S_k \in \{C_1,\ldots,C_R\}$ has N_k refined global centroids in the second layer. In the second layer refined clusters selection, we first rank the distances d_i in ascending order, and then select the top $V\%$ of the centroids as candidate GPS locations for the input image. Thus, the number of selected centroids in the second layer is $F=VN/100$. We denote the selected candidates as $SC=\{g_1,\ldots,g_F\}$ with $g_f \in \{r_1,\ldots,r_N\}(f \in \{1,\ldots,F\})$. In this paper, we user $V=100$ the same as [9].

4.2. Local feature refinement

Local feature matching is utilized to improve GPS location estimation performance. As representative images for each of the second-layer clusters have already been selected in the offline system, we carry out local feature matching for the input image with the representative images of the selected candidates $SC=\{g_1,\ldots,g_F\}$. In this paper we use the inverted file structure based approach to carry out local feature refinement [9].

For each BoW in the input image, we use the obtained inverted files to compute the matching scores of the BoW to the images in the selected candidates $SC=\{g_1,\ldots,g_F\}$. The score is computed based on Term Frequency-Inverse Document Frequency (TF-IDF). The score of the representative $image_L$ to the input image (denoted as $Score(L)$) is assigned as the sum of the scores of all the BoW. The score of each image is computed as follows:

$$Score(L)=\sum_{x=1}^{Q}\frac{W_x * Freq_L(x)}{Number_L * Frequen_x} \tag{1}$$

where $Freq_L(x)$ is the frequency of BoW_x and $Number_L$ is the number of BoW in $image_L$. Frequenx is the frequency of BoW_x in the whole dataset. W_x is the weight of BoW_x in the input image, computed as

$$W_x=\frac{Freq_{input}(x)}{Number_{input}} \tag{2}$$

where $Freq_{input}(x)$ denotes the frequency of BoW_x and $Number_{input}$ is the number of BoW in the input image.

4.3. Neighbor image groups selection

The score of each image cluster can be utilized to rank the final result and to estimate the GPS location. For example, in our previous work, a K-NN based approach is utilized in GPS estimation for the input image. It is likely that images taken from a certain place can be distributed into different clusters due to the differing appearances of the images taken at different times and viewpoints. Thus, K-NN is necessary for improving the GPS location estimation performance. By a simple verification approach, the GPS estimation performance is satisfactory. However, the K-NN based approach is a majority takes all approaches. It only takes the coarse image group information into account but does not make full use of the local saliency of input image. It overlooks the input images relation with the selected second layer image groups, which can be used

to mine out the salient parts. Our previous approach [9] can be viewed as treating the contributions of all local salient parts of the input image identical in its GPS estimation. In this paper, instead of utilizing K-Nearest Neighbor (K-NN) [9] to estimate the geo-location for the input image, we select k nearest image groups, denoted by $SC=\{g_1,\ldots,g_k\}$, to carry out the following refined estimation. The algorithm is as follows:

1. Initialize $SC=\varnothing$ (without element in it), and iteration times $t=1$, and rank the score $Score(L)$ for all images.
2. Select the image group g_1 of the nearest image as the neighbor, update $SC=\{g_1\}$ and $t\leftarrow t+1$.
3. Determine the group information of the tth image in the top ranked image list, if its group information g_t is not in SC, then select it as neighbor, otherwise go to check the next image.
4. Iteratively carry out step (3), until the image group number in SC reaches k, i.e. $SC=\{g_1,\ldots,g_k\}$.

4.4. Salient feature mining

In our previous GPS estimation approach [9], the saliency of local parts of input image was considered identical. That is to say the weights of all the BoWs were the same during similarity measurement. In this paper, saliency of different parts is fully taken into consideration in BoW weight computation.

For the input image I, after SIFT feature extraction and hierarchical quantization, we represent the image as a set of BoW. We denote $I=\{W_1,\ldots,W_N\}$. W_i is the corresponding BoW of the ith SIFT feature, and N is the total number of BoW in the image I.

Fig. 2 shows our approach intuitively. It shows three candidate neighboring image groups. Only the image group at the left side which marked out with the red frame is relevant to the input image, the other two image groups are irrelevant to the input image as shown on its top and at right side. We find that even though the erroneous image groups contain more SIFT correspondences, their distribution is diverse. Conversely, the left image group contains fewer points but there are points which have very high frequency. It is reasonable to assign larger weights for these salient points. Thus by adding weights to the matched SIFT points, we can get their contributions in image GPS localization. As shown in Fig. 2, the red points are the corresponding SIFT feature points, and the corresponding sizes of the points represent their contribution. From Fig. 2, we find that the salient parts on the building of the image assigned high weights, while these in the tree have been assigned low weights.

Our salient feature mining based approach is like the traditional TF-IDF, but with some differences. TF-IDF based approach can be viewed as a globally constrained approach for all SIFT points. TF-IDF has a shortcoming that it cannot give a high weight to the BoW which has strong descriptive ability in some limited classes. For example, some BoWs appear frequently in the certain image groups but very rarely in the total image dataset. While in our salient feature mining approach, we determine the saliency of BoWs of the input image by taking its neighboring image groups into account. Instead of using IDF directly, we also take the social community information into consideration. Let w_j denote the saliency of jth word W_j. We determine the weight as follows:

$$W_j=\frac{n_j}{\sum_i n_i}*\log\frac{|SC|*|C|}{\{i:W_j\in C_i\}*\{i:W_j\in sg_i\}} \tag{3}$$

where n_j is the frequency of BoW W_j in image I, $\sum_i n_i$ stands for the number of BoW in image I, $|SC|$ is the image number in SC and $|C|$ is the number of image groups in SC, $\{i:W_j\in C_i\}$ stands for the number of image groups which contain BoW W_j and $\{i:W_j\in sg_i\}$ is the number of images which contain BoW W_j. As $SC=\{g_1,\ldots,g_k\}$, there are totally k selected neighbors, $|C|=k$. The relationship of this parameter k with the image GPS location estimation performance is discussed in Section 5.4.

4.5. Inverted file referring and scores calculating

Note that image I is represented by a set of BoWs, i.e. $\{W_1,\ldots,W_N\}$, and we get their weights $\{w_1,\ldots,w_N\}$. Thus, for an image L in the dataset, its similarity score to the input image, denoted by $SS(L)$, can be determined from saliency constrained approach as follows:

$$SS(L)=\sum_{\{j:W_j\in L\}} e_j*w_j \tag{4}$$

where $\{j:W_j\in L\}$ stands for all the visual words that are contained in image L, and e_j denotes that whether the W_j exists in the input image, thus we have

$$e_j=\begin{cases}1 & \text{if } W_j\in I\\ 0 & \text{otherwise}\end{cases} \tag{5}$$

4.6. GPS estimation using image ranking and K-NN

Based on the scores of images in dataset to the input image, we rank all the images. The GPS of the K nearest neighbor (K-NN) is utilized to estimate the location of the input image. This approach is similar to that utilized in [9]. First, we use mean-shift clustering for the GPS locations of the images of the selected top ranked K representative images. Finally, we pick out the cluster with the highest cardinality and assign its GPS coordinates to the input image.

5. Experiment and analysis

In order to test the performance of the proposed GPS estimation approach, comparisons are made with GPS estimation [1,2,9] and TF-IDF technique and ours.

To testify the effectiveness of our proposed method, we perform our experiments on GOLD [2,9], OxBuild [14], COREL [16]. All experiments are done on a server with 2.0 GHz CPU and 24 GB memory, and all the experiments are performed on the environment of C. The experimental results demonstrate that the proposed method outperforms the state-of-the-art methods.

Fig. 2. An image and its neighboring image groups. The red points denote the SIFT feature points, with the sizes denote their weights.

5.1. Experimental datasets

The categories of **OxBuild5000** and **COREL5000** serve as GPS locations. Thus, the GPS location numbers of OxBuild5000 and COREL5000 are 14 and 50 respectively. Hundred images are selected randomly from the whole dataset as the test set, while the rest serve as the training set of the offline system for construction of the hierarchical structure.

GOLD contains more than 3.3 million geo-tagged images. It has been compiled from Flickr using its public API and crawlers. Eighty travel spots are selected for testing, i.e. the GPS location number is 80. Thirty-four out of the 80 locations are landscape such as parks, squares, and the rest 46 are landmarks. The test dataset for the 80 sites contains 52,046 images [2]. The experiment shows that our method works both with landmark and landscape.

5.2. Performance evaluation

The performance evaluation contains two parts. The first part tests cross validation performances which utilizes images taken outside the GPS locations in offline systems as input. The second part tests the average recognition rate of test images taken from the GPS locations in offline systems.

As for the test images taken from places in the offline systems, if the selected image group is actually the same group as the test image is from, the estimation is correct. Otherwise, it gives an incorrect estimation. We use average recognition rate (AR) to evaluate the GPS estimation performance, which is given as follows:

$$AR = \frac{1}{G}\sum_{i=1}^{G} A_i \tag{6}$$

where A_i is the correct recognition rate of the ith spot

$$A_i = \frac{NC_i}{NA_i} * 100\%, \quad i \in \{1, \ldots, G\} \tag{7}$$

where NC_i is the correct estimated image number, and NA_i is the test image number. G is the total number of GPS locations.

5.3. GPS estimation performance comparisons

For fair comparison, only visual features of the input image are utilized. As for GPS Estimation (denoted as MSD) [2], we set the parameters of best performance. In the traditional TF-IDF (denoted as TF-IDF) technique, the scale of visual words is 68,420, which is the same as our proposed method (denoted as SILE). Spatial coding based approach (denoted as SC) [18], SVM based landmark classification method (denoted as LC) [23]. As for SC, K-NN is utilized and K is set to be 120 to achieve best performance. As for LC, it mentions that it performs better as the size of codebook increases, so the codebook is set to 68,420. The time cost of LC is computed by only considering the test time without considering the training time. The parameters in our baseline algorithm are set as follows: the social community number $k=20$, the final K nearest neighbors $K=50$, and the size of BoW is set to 68,420. The performance of our approach under the method of K-NN is evaluated. The GPS location estimation performance of MSD, SC, LC, TF-IDF, IFS [9] and SILE are shown in Table 1. The corresponding computational costs are shown in Table 2. For more comparisons, we also provide the performances of SC, LC and other relevant approaches in the following tables and figures.

It can be observed that our method achieves significantly better performance than the other methods not only in GOLD but also in both OxBuild5000 and COREL5000. The average precisions of MSD on the three

Table 1
Average recognition rate(%) of IM2GPS, MSD, SC, LC, TF-IDF,IFS and SILE on COREL5000, Oxbuild5000 and GOLD.

Dataset	IM2GPS	MSD	SC	LC	TF-IDF	IFS	SILE
COREL5000	45.98	97.00	76.01	49.43	48.00	91.00	98.00
Oxbuild5000	39.67	90.00	60.87	53.94	42.10	87.00	92.36
GOLD	53.06	85.02	71.84	54.25	34.29	83.94	89.34

Table 2
Average computational cost (in ms) of IM2GPS, MSD, SC, LC, TF-IDF, IFS and SILE on COREL5000, Oxbuild5000 and GOLD.

Dataset	IM2GPS	MSD	SC	LC	TF-IDF	IFS	SILE
COREL5000	60.46	0.82	7.94	1.04	0.09	0.07	0.10
Oxbuild5000	33.74	0.50	5.42	1.34	0.13	0.09	0.15
GOLD	64927	1.03	47.00	2.89	0.84	0.16	0.56

test dataset are 97%, 90% and 85.02%. The average precisions of TF-IDF are 48%, 42.1% and 34.29%. Those of ours for the three datasets are 98%, 92.36%, and 89.34%. Our method achieves about 3% improvement over MSD and nearly 125% improvement over TF-IDF on average. In our method, the consideration of social information by using social based visual words weight computation really improves performance. Although MSD combines the local feature and local feature together, it totally neglects the social information, which could be the reason for the difference in performance. There are two reasons for the relatively low recognition rates for TF-IDF. One is that spatial information is somewhat neglected while using the BoW histogram. The other is that global features and social information are not taken into consideration. Although SC utilizes local features, it neglects the clues that global features can provide. Thus, our method achieves better performance. There are two reasons for the relatively low recognition rates of LC. One is that spatial information is somewhat neglected using the BoW histogram. The other is that SVM classifiers are affected by the outliers (images with incorrect GPS information) in training. The performance of our method benefits from two facts, through social based visual word weight computation which strengthens the weight of important words and through noisy word suppression, which means the visual words which have little to do with the GPS location estimation are effectively restrained.

The average computational costs of MSD on the three test sets are 0.82 milli-second (ms), 0.5 ms and 1.03 ms, while that of TF-IDF are 0.09 ms, 0.13 ms and 0.84 ms on the three test sets. The average computational costs of *SC* are 7.30 ms, 5.51 ms and 39.60 ms on the three test sets. The *LC* is time efficient with its computational costs 1.04 ms, 1.34 ms and 2.89 ms. The SILE is also time efficient with its computational costs 0.10 ms, 0.15 ms and 0.56 ms. Comparatively, both TF-IDF and SILE are much more time efficient than MSD. Compared with TF-IDF, our method consumes a little more time in dataset of COREL5000 and OxBuild 5000, however, in GOLD, SILE shows its efficiency. As GOLD is a much larger dataset, *SILE* shows its potential in large scale image datasets.

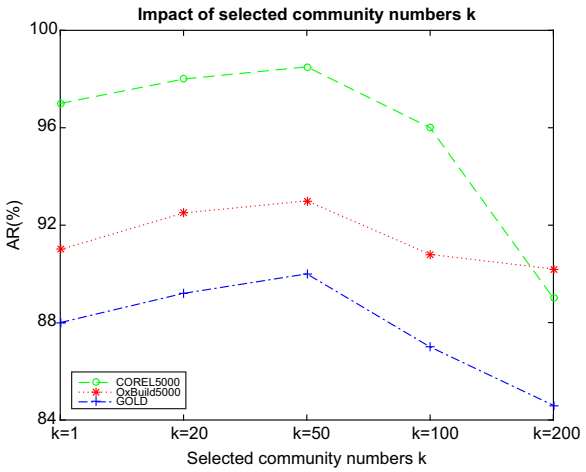

Fig. 3. Plot for analyzing impact of *k*.

It gets highest GPS estimation accuracy among the compared approaches on all the three datasets. At the same time the computational costs are lower than MSD even if they are a bit higher than IFS.

5.4. Discussions

The performance of our approach is related to two parameters: the number k of social communities selected in the first step, and K in K-NN in the last step. The parameters in our baseline algorithm are set as $k=20$, and $K=50$. We will next examine their respective impacts by carrying out a set of experiments on Corel5000, OxBuild5000 and GOLD. For the other parameters, the detailed discussions are given in[9], thus we just focus on the discussions of two introduced parameters k and K

- Impact of total number of selected social communities-k
  To study the impact of the number of selected social communities, we carry out experiments under different k by fixing $K=50$. As shown in Fig. 3, with the increase in k, GPS location estimation accuracy first increases and then decreases. This is due to the fact that, with the increase of number of social communities, it will start to introduce many irrelevant image groups taken from other GPS locations. Accordingly, the possibility of giving a high weight to some insignificant visual word in the input image will increase. Comparatively better performances are achieved under $k=50$.
- Impact of K in K-NN
  To study the impact of K in K-NN on GPS estimation performances at the last step, we carry out experiments on the three datasets under different K by fixing the selected social community number $k=20$. Accordingly, the performances under $K=1$, 10, 50, 100 and $K=1000$ are shown in Fig. 4. We find that when K is set to a very large value, such as $K=1000$, the final performances are very unsatisfactory. This is caused by the introduction of many irrelevant GPS locations. We also find that the GPS estimation performances on GOLD are very stable when K is in the range $[1, 100]$.

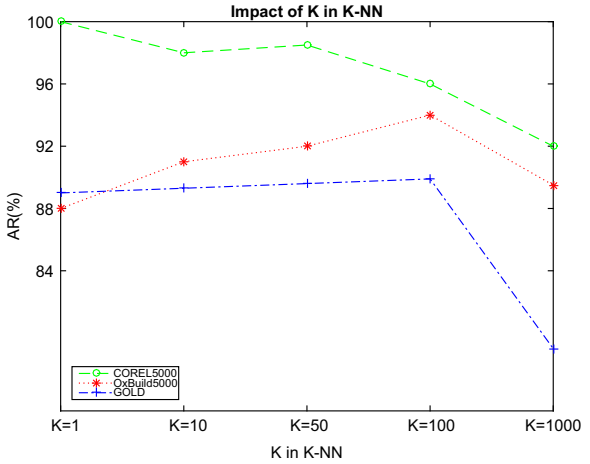

Fig. 4. Plot for analyzing impact of *K*.

6. Conclusion

In this paper, we integrate the social information and image content to accomplish the task of location estimation. We present a salient feature mining based approach to improve the performance of GPS location estimation. We use the salient part of an image by utilizing bag-of-words model of low level SIFT features. Thus different parts of an image should have different contributions to its location estimation. Assigning each BoW with corresponding salient region contributes to the input image location estimation. Experiments demonstrate that our method outperforms the classical and state-of-the-art method for image GPS estimation.

References

[1] J. Hays, A.A. Efros, im2gps: estimating geographic information from a single image, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2008.
[2] J. Li, X. Qian, Y. Tang, L. Yang, C. Liu, Gps estimation from users photos, in: Advances in Multimedia Modeling, Lecture Notes in Computer Science, vol. 7732, Springer, Berlin, Heidelberg, 2013, pp. 118–129.
[3] M. Trevisiol, J. Delhumeau, H. Jégou, G. Gravier, How inria/irisa identifies geographic location of a video, in: Working Notes Proceedings of the MediaEval 2012 Workshop, 2012.
[4] L.T. Li, J. Almeida, D.C.G. Pedronette, O.A.B. Penatti, R. da Silva Torres, A multimodal approach for video geocoding, in: MediaEval, 2012.
[5] X. Li, C. Hauff, M. Larson, A. Hanjalic, Preliminary exploration of the use of geographical information for content-based geo-tagging of social video, in: MediaEval, 2012.
[6] X. Qian, X. Liu, C. Zheng, Y. Du, X. Hou, Tagging photos using users' vocabularies, Neurocomputing 111 (2013) 144–153.
[7] S. Jiang, X. Qian, Y. Xue, F. Li, X. Hou, Generating representative images for landmark by discovering high frequency shooting locations from community-contributed photos, in: 2013 IEEE International Conference on Multimedia and Expo Workshops (ICMEW), IEEE, 2013, pp. 1–6.
[8] S. Jiang, X. Qian, K. Lan, L. Zhang, T. Mei, Mobile multimedia travelogue generation by exploring geo-locations and image tags, in: 2013 IEEE International Symposium on Circuits and Systems (ISCAS), IEEE, 2013, pp. 881–884.
[9] J. Li, X. Qian, Y.Y. Tang, L. Yang, T. Mei, Gps estimation for places of interest from social users' uploaded photos, IEEE Trans. Multimed. 15 (8) (2013) 2058–2071.
[10] P. Kelm, S. Schmiedeke, T. Sikora, Video2GPS: Geotagging Using Collaborative Systems, Textual and Visual Features: MediaEval 2010 Placing Task.
[11] G. Salton, C. Buckley, Term-weighting approaches in automatic text retrieval, Inf. Process. Manag. 24 (5) (1988) 513–523.
[12] Y. Ko, A study of term weighting schemes using class information for text classification, in: Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, 2012, pp. 1029–1030.
[13] J. Sivic, A. Zisserman, Efficient visual search of videos cast as text retrieval, IEEE Trans. Pattern Anal. Mach. Intell. 31 (4) (2009) 591–606.
[14] J. Philbin, O. Chum, M. Isard, J. Sivic, A. Zisserman, Object retrieval with large vocabularies and fast spatial matching, in: Computer Vision and Pattern Recognition, 2007, IEEE, 2007, pp. 1–8.
[15] Y. Li, B. Geng, Z.-j. Zha, Y. Li, D. Tao, C. Xu, Query expansion by spatial co-occurrence for image retrieval, in: Proceedings of the 19th ACM International Conference on Multimedia, ACM, 2011, pp. 1177–1180.
[16] P. Duygulu, K. Barnard, J.F. de Freitas, D.A. Forsyth, Object recognition as machine translation: learning a lexicon for a fixed image vocabulary, Computer Vision ECCV 2002, Springer, 2002, 97–112.
[17] W. Zhou, Y. Lu, H. Li, Y. Song, Q. Tian, Spatial coding for large scale partial-duplicate web image search, in: Proceedings of the International Conference on Multimedia, ACM, 2010, pp. 511–520.
[18] S. Zhang, Q. Huang, G. Hua, S. Jiang, W. Gao, Q. Tian, Building contextual visual vocabulary for large-scale image applications, in: Proceedings of the International Conference on Multimedia, ACM, 2010, pp. 501–510.
[19] H. Liu, T. Mei, J. Luo, H. Li, S. Li, Finding perfect rendezvous on the go: accurate mobile visual localization and its applications to routing, in: Proceedings of the 20th ACM International Conference on Multimedia, ACM, 2012, pp. 9–18.
[20] C. Wu, F. Fraundorfer, J.-M. Frahm, M. Pollefeys, 3d model search and pose estimation from single images using vip features, in: Computer Vision and Pattern Recognition Workshops, 2008, IEEE, 2008, pp. 1–8.
[21] G. Baatz, K. Köser, D. Chen, R. Grzeszczuk, M. Pollefeys, Leveraging 3d city models for rotation invariant place-of-interest recognition, Int. J. Comput. Vis. 96 (3) (2012) 315–334.
[22] M. Park, J. Luo, R.T. Collins, Y. Liu, Beyond gps: Determining the camera viewing direction of a geotagged image, in: Proceedings of the International Conference on Multimedia, ACM, 2010, pp. 631–634.
[23] Y. Li, D.J. Crandall, D.P. Huttenlocher, Landmark classification in large-scale image collections, in: 2009 IEEE 12th International Conference on Computer Vision, IEEE, 2009, pp. 1957–1964.
[24] P. Kelm, S. Schmiedeke, J. Choi, G. Friedland, V.N. Ekambaram, K. Ramchandran, T. Sikora, A novel fusion method for integrating multiple modalities and knowledge for multimodal location estimation, in: Proceedings of the 2nd ACM International Workshop on Geotagging and its Applications in Multimedia, ACM, 2013, pp. 7–12.
[25] Y. Gao, M. Wang, Z.-J. Zha, J. Shen, X. Li, X. Wu, Visual–textual joint relevance learning for tag-based social image search, IEEE Trans. Image Process. 22 (1) (2013) 363–376.
[26] D. Kim, S. Rho, E. Hwang, Local feature-based multi-object recognition scheme for surveillance, Eng. Appl. Artif. Intell. 25 (7) (2012) 1373–1380.
[27] D.G. Lowe, Distinctive image features from scale-invariant keypoints, Int. J. Comput. Vis. 60 (2) (2004) 91–110.
[28] X. Qian, X.-S. Hua, P. Chen, L. Ke, Plbp: An effective local binary patterns texture descriptor with pyramid representation, Pattern Recognit. 44 (10) (2011) 2502–2515.
[29] W. Zhang, S. Shan, W. Gao, X. Chen, H. Zhang, Local Gabor binary pattern histogram sequence (lgbphs): a novel non-statistical model for face representation and recognition, in: Tenth IEEE International Conference on Computer Vision, 2005, vol. 1, IEEE, 2005, pp. 786–791.
[30] A. Torralba, K.P. Murphy, W.T. Freeman, M.A. Rubin, Context-based vision system for place and object recognition, in: Proceedings. Ninth IEEE International Conference on Computer Vision, 2003, IEEE, 2003, pp. 273–280.
[31] P. Tirilly, V. Claveau, P. Gros, Language modeling for bag-of-visual words image categorization, in: Proceedings of the 2008 International Conference on Content-Based Image and Video Retrieval, ACM, 2008, pp. 249–258.
[32] K. Kesorn, S. Poslad, An enhanced bag-of-visual word vector space model to represent visual content in athletics images, IEEE Trans. Multimed. 14 (1) (2012) 211–222.
[33] J. Li, X. Qian, Q. Li, Y. Zhao, L. Wang, Y.Y. Tang, Mining near duplicate image groups, Multimed. Tools Appl. 74 (2) (2014) 655–669.

[34] Y. Zhao, X. Qian, T. Mu, Image taken place estimation via geometric constrained spatial layer matching, MultiMedia Modeling, Springer, 2015, 436–446.
[35] X. Yang, X. Qian, Spatial verification for scalable mobile image retrieval, in: Proceedings of the 23rd ACM International Conference on Information and Knowledge Management, ACM, 2014, pp. 1903–1906.
[36] X. Yang, X. Qian, T. Mei, Learning salient visual word for scalable mobile image retrieval, Pattern Recognit.
[37] Y. Zhao, X. Qian, Spatial constraint for image location estimation, in: ICMR, 2015.
[38] X. Yang, X. Qian, Y. Xue, Scalable mobile image retrieval by exploring contextual saliency, IEEE Trans. Image Process. (2015), In press.
[39] Z. Liu, W. Zou, O. le Meur, Saliency tree: a novel saliency detection framework, IEEE Trans. Image Process. 23 (5) (2014) 1937–1952.
[40] J. Knopp, J. Sivic, T. Pajdla, Avoiding confusing features in place recognition, in: Computer Vision—ECCV 2010, Springer, 2010, pp. 748–761.
[41] J. Han, S. He, X. Qian, D. Wang, L. Guo, T. Liu, An object-oriented visual saliency detection framework based on sparse coding representations, IEEE Trans. Circuits Syst. Video Technol. 23 (12) (2013) 2009–2021.
[42] J. Han, D. Zhang, X. Hu, L. Guo, J. Ren, F. Wu, Background prior based salient object detection via deep reconstruction residual, IEEE Trans. Circuits Syst. Video Technol. (2014), In press.
[43] L. Zhu, J. Shen, H. Jin, R. Zhang, L. Xie, Landmark classification with hierarchical multi-modal exemplar feature, IEEE Trans. Multimed.
[44] L. Zhu, J. Shen, H. Jin, R. Zhang, L. Xie, Content-based visual landmark search via multimodal hypergraph learning, IEEE Trans. Cybern.