Spatial Constraint for Image Location Estimation

Yisi Zhao
SMILES LAB, Xi'an Jiaotong University
Xianning West Road, Xi'an, China
zyswhy0203@stu.xjtu.edu.cn

Xueming Qian
SMILES LAB, Xi'an Jiaotong University
Xianning West Road, Xi'an, China
qianxm@mail.xjtu.edu.cn

ABSTRACT
Nowadays, image location has been widely used in many application scenarios for large geo-tagged image corpora. As to images which are not geographically tagged, we can estimate their locations with the help of the large geo-tagged image set by content based image retrieval. In this paper, we propose a global feature clustering and local feature refinement based image location estimation approach. We exploit spatial information by processing useful visual words. In this process, visual word groups are generated. Moreover to improve the retrieval performance, spatial constraint is utilized to code the relative position of visual words. Here we generate a position descriptor for each visual word. Experiments show the effectiveness of our proposed approach.

Categories and Subject Descriptors
H.3.1 [**Information Systems**]: Information storage and retrieval-*content analysis and indexing.*

General Terms
Algorithms, Measurement, Performance, Experimentation

Keywords
Location estimation; Bag-of-words; Visual word group; Position descriptor

1. INTRODUCTION
In recent years, large quantities of images taken by the users are shared in social media websites such as Facebook, and Flickr every day. Many of the images are associated with the locations they were taken. As to images without geo-tags, automatic location estimation for them is possible with the help of the large scale geo-tagged photos. In this paper, our task is to estimate the location of an input image by content based image retrieval approach. State-of-the-art large scale image retrieval systems have relied on the bag-of-words (BoW) model and local descriptors, such as SIFT, SURF. Li et al. utilize multi-class SVM classifiers using bag-of-words for large scale image location estimation [3]. However, there still exist deficiencies in BoW model. Many improved approaches are proposed to enhance the discrimination, e.g. visual synonyms, embed geometry constraint [2]. Moreover, the database can be constructed with a 3D model. Liu et al. proposed an approach to estimate accurate parameters about the scene geo-information [5]. Park el al. proposed a method of viewing direction determination by utilizing Google Street View and Google Earth satellite [7]. In this paper, we further explore global feature clustering and local feature refinement based approach to complete image location estimation.

Experimental results of existing work show that the commonly generated visual words are still not as expressive as the text words. Wu et al. [1] employ Maximally Stable Extremal Regions (MSER) to bundle SIFT features into groups instead of taking all of them individually. Moreover, spatial verification enforces geometric consistent constraint on visual words that query and dataset image share, such as RANSAC and spatial coding [2]. Spatial information of visual words should be exploited for better image retrieval performance. In our work, visual words mining and spatial constraint based image location estimation approach is exploited. Firstly, we determine the refined locations of an input image using global features clustering. This step can speed up the image location estimation process by selecting candidate locations. Considering that the distribution of an image's visual words directly reflects the distribution of the image's main content, secondly, we mine the salient features and exploit spatial information to improve the image location estimation performance. In this process, (1) we utilize term frequency-inverse document frequency (tf-idf) to select visual words with higher weight. (2) we divide an image's useful visual words into multiple groups by Mean-shift clustering. A visual word group is composed of visual words in the corresponding cluster. (3) group based spatial coding is conducted. We generate a position descriptor for visual word.

The main contributions of this paper are summarized as following: 1) useful feature selection is utilized, which eliminates the effect of noisy, unstable and irrelevant features; 2) we bundle useful visual words to generate multi-group, and a group based image retrieval method is proposed; 3) spatial information of visual word group is mined. We describe each visual word's distribution in the group it belongs. The rest of the paper is organized as follows: Firstly, we provide the system overview. Secondly, we give a description on our approach in section3-5. Finally, experiments containing the comparison with the recently popular method and discussions are shown in Section6. In Section7, the conclusion is drawn.

2. SYSTEM OVERVIEW
The system of our proposed approach is shown in **Fig.1**. It consists of offline part and online part. In the offline part, we build a hierarchical index for the dataset images [4]. And we build visual word group (VWG) by Mean-shift clustering. Then each visual word's position descriptor is generated for dataset images. In the online part, (1) refined locations of an input image are pre-selected by cluster selection online. In this process, global feature clustering is utilized [4]. (2) local feature of the input image is in

ICMR'15, June 23–26, 2015, Shanghai, China.

ACM 978-1-4503-3274-3/15/06…$15.00
DOI: http://dx.doi.org/10.1145/2671188.2749327

full use. The VWG building and spatial constraint are conducted. (3) we estimate the location of the input image by VWG based image search.

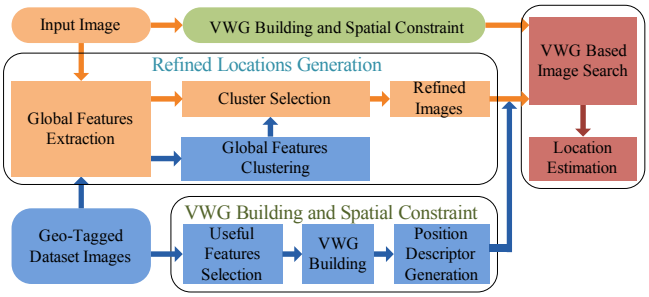


Figure1. Block diagram of the location estimation system.

3. REFINED LOCATIONS GENERATION

In the part, we introduce how to select the refined locations.

We cluster the dataset images using global features. In order to show the effectiveness of our proposed image location estimation approach, we utilize the same visual features and the suggested parameters for global features clustering in [4]. Through global feature clustering, the whole dataset can be divided into several small scale groups. K-means clustering is utilized to divide the dataset into M small clusters, denoted as $C_n(n=1,2,\ldots,M)$. Then, we select candidate clusters according to the distance between the input image and M centers as that in [4]. The top ranked $S(S<M)$ clusters are selected. In this paper, we set S 15. We further obtain occurred locations of images in the selected clusters. The occurred locations are served as the refined locations of the input image.

4. VISUAL WORD GROUP BUILDING AND SPATIAL CONSTRAINT

In this part, we mine visual word groups for each refined image and the input image, and enforce spatial constraint to improve image location estimation performance. The detailed process includes three steps: (1) useful feature selection, (2) visual word group building, (3) position descriptor generation.

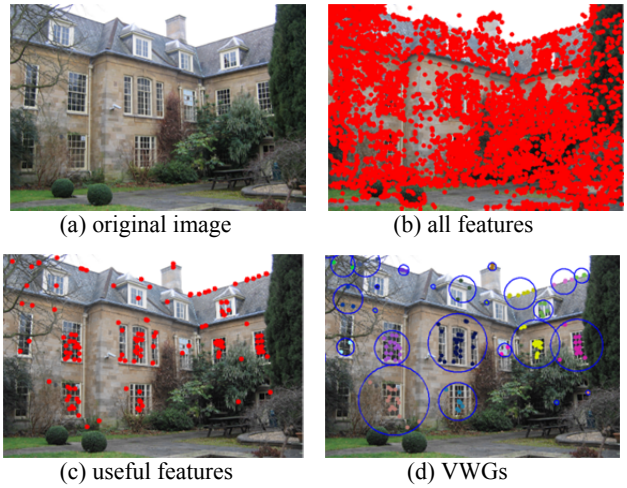
4.1 Useful Feature Selection

For an image as shown in **Fig.2 (a)**, there are 4002 SIFT points detected which are shown in **Fig.2 (b)**. However, different visual words have different weights of importance for identifying the query scene. Some visual words are non-distributive. As shown in **Fig.2 (b)**, many visual words often appear in the part of grass and trees, which are confusing for accurate location estimation. To mine useful features, we compute the score of each word by employing a tf-idf weighting scheme as follows:

$$S_w = \frac{f_w}{\sum_w f_w} \times \log\frac{N}{n_w} \tag{1}$$

where f_w is the frequency of w-th BoW in the image, n_w is the number of images containing the w-th BoW.

We keep the visual words whose scores are larger than *thr*. As shown in **Fig.2(c)**, there are only 169 visual words left, which is far less than the raw SIFT features. In this paper, we set *thr=0.001*. The choice of *thr* does influence the performance, but the impact is small in comparison with its computational cost as shown in our discussion.



(a) original image (b) all features

(c) useful features (d) VWGs

Figure2. For an input image (a), we extract 4002 raw SIFT points which are shown in (b). (c) shows the selected 169 useful visual words. In (d), we generate 24 visual word groups.

4.2 Visual Word Group Building

In this section, we build visual word group (VWG) for each image. We aim at increasing the precision of the traditional bag-of-words representation, because the VWG based methods employ group feature matching instead of single feature matching. For an image, we cluster the coordinates of its useful words by Mean-shift clustering [8]. 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 (x,y) of visual words are utilized. Let $v=\{(x_i,y_i)\}_{i=1}^{h}$ denote the locations of the h SIFT points after useful feature selection. To $\forall v$, Mean-shift is defined as follows:

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

where N_o is the number of observations falling within $S_b(v)$ region. z is the visual words falling within $S_b(v)$ region. b is the bandwidth parameter [8].

After clustering, we obtain several clusters and their corresponding centers. A cluster is considered as a VWG, which is composed of all visual words in the cluster. Until now, we represent an image with multiple VWGs. Assuming that there are L VWGs generated, we denote them as $G_l, l=1,2,\ldots,L$. For the useful words shown in **Fig.2 (c)**, the corresponding VWGs are shown in **Fig.2 (d)**. Totally there are 24 VWGs. In order to visually display the VWGs, we mark a unique color for each VWG.

4.3 Position Descriptor Generation

This section presents our approach to represent the spatial information of visual words. We generate a position descriptor (PD) for each visual word to describe the distribution of a visual word in its corresponding VWG. Assuming that a VWG G_l has

n visual words which are denoted by $\{w_1, w_2, \ldots, w_n\}$), our representation of PD includes the following two aspects.

4.3.1 RA constraint

We set the center of a cluster as the center of the corresponding VWG. We divide the VWG space into quadrants using its center as the origin of the quadrants. For each word w_i in the G_l , we record its spatial position in relation to the origin. It reveals, for example, that a visual word tends to be below, at right relative to the center. For a VWG's word, its position matrix is defined as follows.

$$RA_i = \begin{cases} [1\,0\,0\,0], \text{if } x_i > a_0, y_i > b_0 \\ [0\,1\,0\,0], \text{if } x_i < a_0, y_i > b_0 \\ [0\,0\,1\,0], \text{if } x_i < a_0, y_i < b_0 \\ [0\,0\,0\,1], \text{if } x_i > a_0, y_i < b_0 \end{cases} \quad (3)$$

where (x_i, y_i) is the coordinates of word w_i . (a_0, b_0) are the coordinates of the center of the VWG. Thus the RA of a visual word is a 4 bit vector, which shows its relative spatial area.

4.3.2 RD constraint

We calculate the relative distance (RD) between the visual word and the center of the VWG. We calculate the distance of each visual word w_i and the center by Euclidean distance. We denote d_i as the distance of visual word w_i and the center. The relative distance of word w_i is calculated like this:

$$RD_i = \frac{d_i}{\frac{1}{n}\sum_{i=1}^{n} d_i} \quad (4)$$

We further make the normalization on the value of RD. If a word's RD is less than or equal to one, we think that the word is near to the center. The normalization is modeled as follows. Until now, we obtain each word's RA and RD relative to the VWG's center. We put them together, obtaining a five dimensional vector. The five dimensional vector is the position descriptor of visual word w_i in the VWG, denoted as PD_i .

$$RD_i = \begin{cases} 0, if\ RD_i \le 1 \\ 1, if\ RD_i \ge 1 \end{cases} \quad (5)$$

5. VWG BASED IMAGE SEARCH

An image retrieval method based on the VWG and the spatial geometric consistency is presented in this section. In the online system, the VWGs and the PD vectors of input image are generated. Then we introduce how to calculate the similarity between the input image $\#q$ and the refined image $\#r$. The process includes the following two steps.

The first step is matched group pair (MGP) detection. Let G_i^q denotes the *i*-th VWG from image $\#q$. And G_i^r is denoted as the *j*-th VWG from image $\#r$. We call the two VWGs as a MGP if they contain common visual words. The matching score of each MGP is calculated from their common visual words and their corresponding PDs. Assuming that G_i^q and G_i^r share a visual words, their corresponding PDs are denoted as $PD_q^k, (k = 1, 2 \ldots, a)$ and $PD_r^k, (k = 1, 2 \ldots, a)$ separately. By comparing PD_q^k and PD_r^k , we calculate the MGP's matching score (MS_{MGP}) as the following:

$$MS_{MGP} = \frac{1}{a}\sum_{k=1}^{a} PD_q^k \oplus PD_r^k \quad (6)$$

where $\oplus$ is Logical Exclusive OR (XOR) operation. The smaller MS_{MGP} means that spatial consistent score of the MGP is higher. Thus the two images are more similar. If some parts of two images match well, we can get images we want. It has better performance than that way of considering the entire content of an image. Moreover, our VWG is unbounded with its position and shape.

The second step is MS_{MGP} selection. Assuming that the input image $\#q$ and the refined image $\#r$ have m MGPs, we can obtain m values about MS_{MGP} . In this paper, the minimum value is selected as the score of the refined image to the input image. Then we rank the refined images based on their scores. Moreover, we take the number of MGPs into consideration. The initial results are re-ranked according to the number of MGPs. At last, the top ranked k images are selected. We count the number of images for each occurred location. The majority location in the k images is assigned for the input image.

6. EXPERIMENTATION

In order to test the performance of the proposed location estimation approach, comparisons are made with IM2GPS [9], CS [4], spatial coding based approach [2] (denoted as SC), method of salient region mining using maximally stable extremal region [1] (denoted as MSER), method of adopting word spatial arrangement [6] (denoted as WSA) and ours (denoted as VWG). Here, the method MSER is that we utilize maximally stable extremal region [1] instead of our VWG to bundle visual words into groups. Other parts of method MSER is the same as our method. We do this is to show the effectiveness of our mean-shift based salient group detection. Similarly, the method WSA is that we adopt word spatial arrangement (WSA) [6] instead of position descriptor generation in our method.

6.1 Datasets

Experiments are done on two datasets: OxBuild, GOLD. OxBuild is used for preliminary tests. The GPS numbers of OxBuild5000 is 11. 100 images are selected randomly from the whole dataset 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 [4]. And 80 travel spots are selected, i.e. the number of locations is 80. The test dataset for the 80 sites contains 5000 images.

6.2 Performance Evaluation

For an input image, if the estimated location is exact with ground-truth, it is correctly estimated. Assuming that the recognition rate of the *i*-th spot (RR_i) is the correct, then average recognition rate (AR) is utilized to evaluate the performance.

$$AR = \frac{1}{G}\sum_{i=1}^{G} RR_i \quad (7)$$

$$RR_i = \frac{NC_i}{NI_i} \times 100\%, i \in \{1, 2, \ldots, G\} \quad (8)$$

where NC_i is the correct estimated image number, NI_i is the test image number. G is the number of dataset locations.

6.3 Performance Comparison

The location estimation performance of IM2GPS, SC, CS, MSER, WSA and VWG are shown in **Fig.3**. It is clear from **Fig.3** that our method can outperform the other methods in OxBuild and GOLD. The average recognition rates of spatial coding (SC) in the two test dataset are 70.39% and 59.48% while the results of Cosine Similarity (CS) are 89.27% and 84.86%. The performances of MSER in the two test dataset are 91.12% and 85.47%. The performances of our VWG in the two test dataset are 93.85% and 88.16% which get performance improvements.

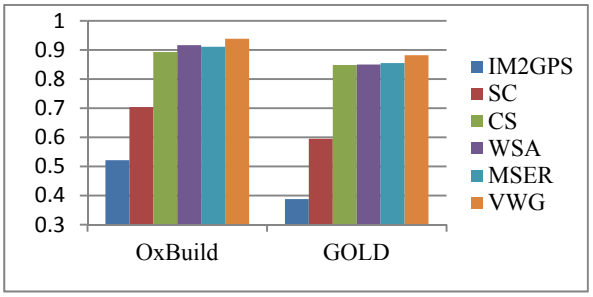


Figure3. ARs of IM2GPS, SC, CS, WSA and MSER, VWG.

6.4 Discussion

The performance of our approach is influenced by several factors. Hereinafter, we carry out the discussion.

6.4.1 The impact of using useful features or not

For images, some visual words may be semantically closer to a certain scene. We carry out the useful feature selection. The situation that all features are used for image retrieval is discussed here. It can be seen from **Table1** that the performance of using useful features is larger on the contrary. **Table2** shows the average time cost of using useful features is lower. Selecting salient words is of significance for image retrieval.

Table1. Average recognition rates (%) of using all features and useful features of images

Dataset	All features	Useful features
OxBuild	92.54	93.85
GOLD	86.97	88.16

Table2. The comparison of average computational costs (ms)

Dataset	All features	Useful features
OxBuild	795.201	380.618
GOLD	1017.937	623.143

Table3. Average recognition rates (%) of using RA or RD

Dataset	RA	RD	PD
OxBuild	91.52	90.65	93.85
GOLD	84.70	85.17	88.16

6.4.2 The impact of different spatial contraints

In our experiments, we generate a position descriptor PD for each visual word, which includes two aspects: the relative area RA and the relative distance RD. Here we discuss the impact of using RA or RD respectively to image location estimation performance. The corresponding results are shown in **Table3** respectively. We find that combining both RA and RD better performances are achieved.

6.4.3 The impact of bandwidth b

In the section of VWGs generation, we cluster the useful words by Mean-shift cluster. So, the multi-VWG generation is closely connected with the bandwidth b. Here, we discuss the impact of bandwidth b. **Fig.4** shows that with the increase of b, the AR is first increasing and then into decline. b is set 70 in the baseline.

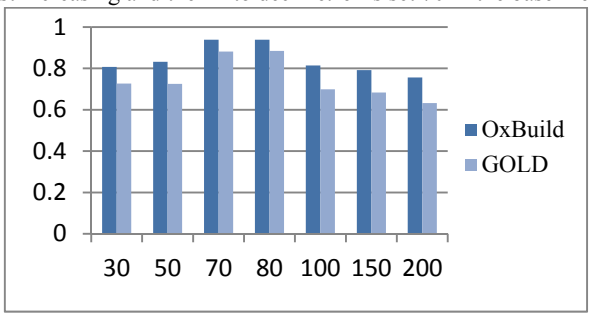


Figure4. Impact of bandwidth b to image location estimation performance.

7. CONCLUSION

In this paper, we present a method of image location estimation, which is based on global feature clustering and local feature refinement. Firstly, refined locations of an input image are pre-selected. Secondly, we localize the image by local feature refinement. In this process, visual word groups are generated and spatial information for words in VWG are coded. At last, group based image search is conducted. Experiments show the effectiveness of our proposed approach.

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

8. REFERENCES

[1] Z. Wu, Q. Ke, M. Isard, and J. Sun, "Bundling features for large scale partial-duplicate web image search," IEEE Conference on CVPR, pp. 25-32, 2009.

[2] W. Zhou, Y. Lu, H. Li, Y. Song, Q. Tian, "Spatial Coding for Large Scale Partial-Duplicate Web Image Search," MM'10, Firenze, Italy. Copyright 2010 ACM.

[3] Y. Li, D.J. Crandall, D.P. Huttenlocher, "Landmark Classification in Large-scale Image Collections," ICCV'09.

[4] Jing Li, Xueming Qian, Yuan Yan Tang, Linjun Yang, and Tao Mei, "GPS estimation for places of interest from social users' uploaded photos", IEEE Trans. Multimedia 2013.

[5] 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," ACM Multimedia, 2012.

[6] Otávio A.B. Penatti, Fernanda B.Silva, Eduardo Valle, "Visual word spatial arrangement for image retrieval and classification," Pattern Recognition 47(2), 2014.

[7] M. Park, J. Luo, R. T. Collins, Y. Liu, "Beyond GPS: Determining the Camera Viewing Direction of A Geo-tagged Image," MM'10, October 25–29, 2010.

[8] K. Fukunaga, L.Hostetler, "The estimation of the gradient of a density function, with applications in pattern recognition," IEEE Transactions on Information Theory, vo1.21, 1975.

[9] J. Hays and A. A. Efros, "IM2GPS: estimating geographic information from a single image," In CVPR, 2008.