Image Taken Place Estimation via Geometric Constrained Spatial Layer Matching

Yisi Zhao, Xueming Qian*, and Tingting Mu

SMILES LAB, Xi'an Jiaotong University, China
zyswhy0203@stu.xjtu.edu.cn, qianxm@mail.xjtu.edu.cn,
T.Mu@liverpool.ac.uk

Abstract. In recent years, estimating the locations of images has received a lot of attention, which plays a role in application scenarios for large geo-tagged image corpora. So, as to images which are not geographically tagged, we could estimate their locations with the help of the large geo-tagged image set by visual mining based approach. In this paper, we propose a global feature clustering and local feature refinement based image location estimation approach. Firstly, global feature clustering is utilized. We further treat each cluster as a single observation. Next we mine the relationship of each image cluster and locations offline. By cluster selection online, several refined locations likely to be related to an input image are pre-selected. Secondly, we localize the input image by local feature matching which utilizes the "SIFT" descriptor extracted from the refined images. In this process, "spatial layers of visual word" (SLW) is built as an extension of the unorganized bag-of-words image representation. Experiments show the effectiveness of our proposed approach.

Keywords: Location Estimation, Spatial Layer Matching, Bag-of-Words.

1 Introduction

Given a query image, in this paper, our goal is to estimate its location by mining image content. Automatic location estimation for an image is possible with the help of the large scale geo-tagged photos shared by millions of worldwide users. State-of-the-art large scale image retrieval systems have relied on local SIFT descriptors [5]. Traditionally, a visual vocabulary is trained by clustering a large number of local feature descriptors. The exemplar descriptor of each cluster is called a visual word, which is then indexed by an integer. However, experimental results of existing work show that the commonly generated visual words are still not as expressive as the text words. Spatial information of visual words should be exploited for better performance. Moreover, we find that although purely using global features is not so efficient, some images can be recognized well via global feature matching.

Therefore, we propose image visual mining based image geographic location estimation approach. In our work, firstly, the clusters are mined to generate refined locations for an input image using global features. Secondly, we exploit sufficient

* Corresponding author.

X. He et al. (Eds.): MMM 2015, Part II, LNCS 8936, pp. 436–446, 2015.

information by mining spatial information of visual words. "Spatial layers of visual word" (SLW) is proposed, which plays a significant role for image location estimation. SLW is generated by involving one visual word and its spatial relationships with its neighbor visual words. Unlike what is introduced in [11], their "spatial pyramid" is generated by partitioning an image into increasingly fine sub-regions and computing histograms of local features found inside each sub-region. We go deep into each word whose multiple neighbors are taken into consideration in sequence.

The contributions of this paper are as follows: (1) Refined locations of an input image are generated via cluster selection based on cluster location estimation. (2) Spatial layer matching is proposed to improve the estimation accuracy for an input image. (3) Useful local features selection is utilized, on the basis of which our proposed SLW shows better results.

The rest of the paper is organized as follows: Firstly, related works on location estimation are reviewed. Secondly, we provide the system overview. Finally, we give a description on our approach in section 4 and 5. Experiments containing the comparison with the recently popular method and parameters discussions are shown in Section 6. In Section 7, the conclusion is drawn.

2 Related Work

Many methods are intended to estimate the geographic location of images. An approach which is based purely on visual features is presented by Hays and Efros in [9]. They characterize each image using a number of image features. Then they compute the distances on different feature spaces and use the k-nearest-neighbor technique to estimate the GPS of an input image. Finally, cluster with the highest cardinality is selected and its GPS is assigned as GPS of the input image

Bag-of-words image representation has been utilized for many multimedia and vision problems. Li et al. utilize multi-class SVM classifiers using bag-of-words for large scale image location estimation [2]. They also show that through adding textual features such as tags, they can improve the performance. Han et al. propose an object-based image retrieval algorithm. They combine a novel feature descriptor based on context-preserving bag-of-words and a two-stage re-ranking technique to measure the similarity between the query image and each image in the dataset [16]. Zhang et al. propose a spatial coding based image retrieval approach by building the contextual visual vocabulary [1]. The spatial coding encodes the relative positions between each pair of features in an image. They focus on user traces across the micro-blogging platform Twitter. Chum et al. also propose an approach for estimating the location of the image by using local feature matching [10]. And user interaction is required to confine the locations of the input image to really small ranges. In [17,18], the GPS information is served as an important clue to improve tag recommendation performances for social user shared photos.

Researchers have proposed many works e.g. visual synonyms [7, 14-15], embed geometry constraint [3, 12-13], etc. Spatial information can reinforce the discriminative power of single word. Wu et al. [12] employed the detector of Maximally Stable Extremal Regions (MSER) to bundle point features (SIFT) into groups instead of taking all of them individually. Moreover, the database can be

constructed with a 3D model [6]. Liu et al. propose an approach which is capable of providing a complete set of more accurate parameters about the scene geo—including the actual locations of both the mobile user and perhaps more importantly the captured scene along with the viewing direction [6]. They firstly perform joint geo-visual clustering in the cloud to generate scene clusters, with each scene represented by a 3D model. The 3D scene models are then indexed using a visual vocabulary tree structure.

3 System Overview

The system of our proposed approach is shown in Figure 1. It consists of two systems: the online system and the offline system.

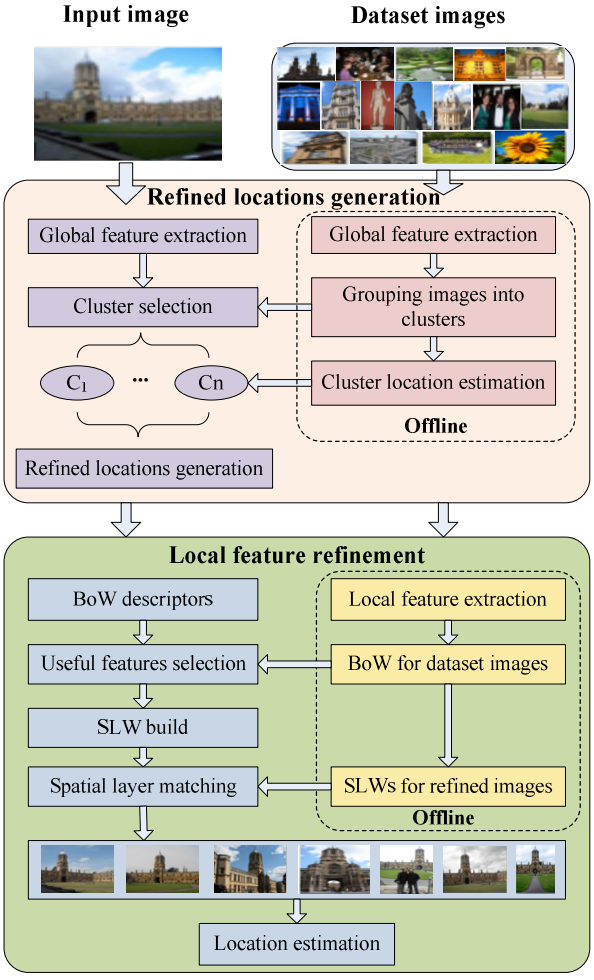


Fig. 1. Block diagram of the location estimation system

Firstly, we obtain refined locations related to an input image offline. In this process, global feature clustering is utilized. We further treat each cluster as a single observation to mine the relationship of each image cluster and locations. Then several refined locations likely to be related to an input image are pre-selected by cluster selection in our online system. Secondly, we estimate image GPS by local feature refinement by making full use of the images in refined locations. In our work, "spatial layers of word" (SLW) is proposed as an extension of bag-of-words image representation. SLWs for dataset images are built offline. We estimate the location of an input image by spatial layer matching.

4 Refined Locations Generation

Generating refined locations is the first step of our framework in our offline system. In this section, we introduce how to generate image clusters, and how to select the refined location candidates.

4.1 Grouping Images into Clusters

We propose to cluster the dataset images using their global features, such as color feature and texture feature. Similar to our previous work [4], color moment (CM) and hierarchical wavelet packet descriptor (HWVP) [19] are utilized here. The global feature clustering is carried out on the 215d vector including 45d CM and 170d HWVP. K-means clustering is utilized to divide dataset images into M clusters $C_i\left(i=1,\ldots,M\right)$. In this paper, we set M to be 50, according to the suggestions in [4].

4.2 Cluster Location Estimation

Due to the fact that our dataset images are geo-tagged, each image has one geo-tag. The geo-tag indicates the taken place of the input image. In our offline system, before the cluster selection, we first mine the relationship of clusters and locations. Our approach consists of the following steps:

Assume that the cluster C_n has g images $I_{nj}\left(j=1,\ldots,g\right)$. Firstly, for each image in the cluster C_n, we gather its R most similar images across the entire dataset images based on the similarities of the global visual features (We will discuss the situation that using local features instead of global features in our experiments.). We select the top ranked $K=R\times g$ neighboring images $I_i\left(i=1,\ldots,\mathrm{K}\right)$.

Secondly, through analysis of the K geo-tagged neighboring images, we can predict the probable locations for the cluster C_n. We divide the geo-tags of the K images into L ($L\leq K$) sets according to their true locations. Each set corresponds to a unique location. Let $D_i\left(i=1,\ldots,L\right)$ denote the L sets. By ranking the L locations according their frequencies (i.e. the numbers of images belonging to the locations) in

the descending orders, we can get the probabilities that the cluster C_n belonging to. As shown in Figure 2, among the K neighbor images, $I_1, I_2, I_3, I_4, I_6, I_{K-1}$ belong to the same location D_1. So, the score (frequency) of location D_1 is 6. We select $V\%$ of the L locations as location candidates related to the cluster.

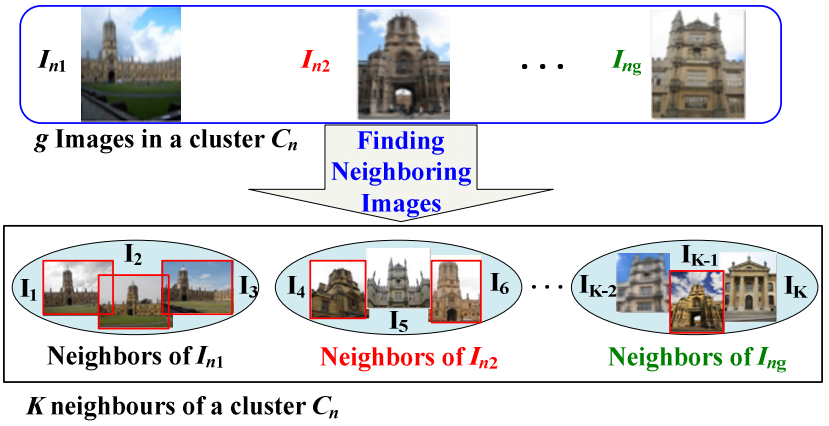


Fig. 2. Finding neighboring images for a cluster

Finally, we select candidate clusters for an input image in our online system. Let F_x denote the 215d global features of the input image. The candidate cluster selection is based on the distances between F_x and M centers $C_i\,(i = 1,\ldots,M)$. In this paper, the top ranked fifteen clusters are selected, i.e. g=15. Based on the found neighboring images for each cluster we can get the refined locations.

5 Local Feature Refinement

After the global feature clustering and refined location generation, we can determine the candidate locations for the input image. In order to improve image location estimation performances, we further conduct local feature refinement. In this section, to capture some unique and representative details in images, we utilize SIFT to carry out spatial layer matching. In our work, we first quantize the SIFT points into visual words by using a hierarchical K-means clustering approach [4].

5.1 Useful Features Selection

Given a query image, its visual words have different discrimination power for location estimation. Some of them are useful for location estimation, and some of them may be noise. To mine useful features, we compute the score of the visual word while considering the frequency and the weight of word by employing a

term-frequency inverse-document-frequency (tf-idf) weighting scheme. For an image, the score of each visual word is computed as follows:

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

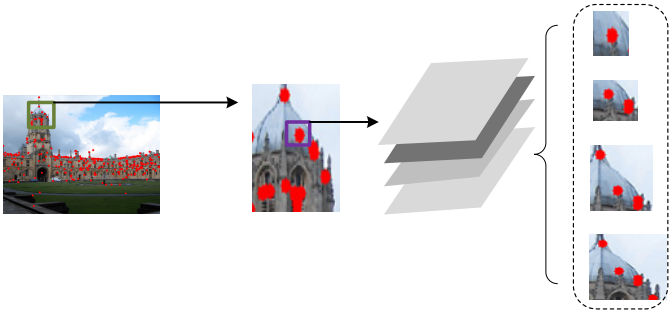
where f_w is the frequency of w-th visual word in the image, n_w is the number of images containing the w-th visual word. In Figure 3 (a) the raw SIFT points are shown, and in Figure 3 (b) the useful feature are kept. We find that by useful feature mining many non-discriminative visual words are removed.

(a) All features (b) useful features

Fig. 3. (a)All features of an image and (b) the useful features

5.2 Spatial Layer Matching

After the above-mentioned steps, each refined image is represented by a set of useful visual words. In this section, we build SLW for each useful visual word of the refined images, which is generated by integrating a visual word and its neighboring visual words. In our feature extraction, we represent each SIFT point by a 128-D descriptor vector and a 4-dimensional DoG key-point detector vector (x, y, scale, and orientation). In this part, the coordinates (x, y) are utilized to calculate the distance of visual words, according to which we build SLW.



(a) (b) (c)

Fig. 4. Spatial layers of visual word (SLW), (a) local region, (b) the neighboring visual words around a visual word, (c) spatial layer representation for a visual word

Therefore, from a visual word w, we build its spatial layers as shown in Figure 4. The enlarged location region for Figure 4 (a) is as shown in Figure 4(b), its spatial layer

representation for the visual word is shown in Figure 4(c). Let $SLW(w)$ to record both the word and its neighbor visual words. We define each layer of $SLW(w)$ as:

$$SLW_n(w) = \left\{ w, (NW_n)_{n=0}^{k} \right\} \tag{2}$$

where NW is neighbors of word w, and n is one of the neighbors of w. k is the number of layers. $SLW_0(w)$ denotes the first layer of w, which only contains the visual word itself. The second layer is composed of the visual word and its nearest neighbor. The third layer consists of the visual word, and its two nearest neighbors, and so on. During our experiments, we build 3 layers for each visual word. The value of n will be discussed in section 6.3.

Then, we score for each image of refined locations like this: For each visual word q_m of a query image, we build its $SLW(q_m)$. For a refined image r, SLW of its any useful visual word g is denoted as $SLW(g)$. If $SLW_n(q_m)$ is found in $SLW(g)$, the score of refined image r (denoted as $Score_r$) is accumulated by one. We assume that query image has Z visual words. Then, we iterate over all the visual words of the query image to calculate the score of the image as follows:

$$Score_r = \sum_{m=1}^{Z} f_m \tag{3}$$

where f_m is an indicator, it records whether the visual word q_m belonging to image r.

$$f_m = \begin{cases} 1, if\ SLW_n(q_m) \in SLW(g),\ g \in r \\ 0, if\ SLW_n(q_m) \notin SLW(g),\ g \in r \end{cases}, n = 0, 1, \ldots, k \tag{4}$$

So, we obtain scores of all candidate images. Then we rank all the candidate images according to their scores. At last, we use K-NN based approach to estimate the location of input image.

6 Experimentation

In order to test the performance of the proposed GPS estimation approach, comparisons are made with IM2GPS [9], CS [4] and spatial coding based approach (denoted as SC) [8]. Experiments are carried out on two datasets: OxBuild and GOLD [4]. The location numbers of OxBuild is 11. 100 images are selected randomly from the whole dataset as the test set, while the rest is served as training set. GOLD contains more than 3.3 million images together with their geo-tags. 80 travel spots are randomly selected for testing. The test dataset for the 80 sites contains 5000 images.

6.1 Performance Evaluation

For an input image, if the estimated location is exact with its ground-truth location, it is correctly estimated, otherwise falsely 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 which is given as follows:

$$AR = \frac{1}{G} \sum_{i=1}^{G} RR_i \tag{5}$$

$$RR_i = \frac{NC_i}{NI_i} \times 100\%, i \in \{1, \ldots, \mathrm{G}\} \tag{6}$$

where NC_i is the correct estimated image number, NI_i is the test image number. G is the number of locations, 11 and 80 for OxBuild and GOLD respectively.

6.2 Performance Comparison

As for IM2GPS, Spatial Coding (SC) and Cosine Similarity (CS), we choose the best parameters provided in [9], [8] and [4]. From Figure 5 we find that our method SLW outperforms the other methods. The results of IM2GPS in the two test datasets are 39.67% and 53.06%. The results of spatial coding (SC) in the two test datasets are 59.48% and 70.39%, while the results of Cosine Similarity (CS) in the two test datasets are 89.27% and 84.86% respectively. Those of ours for the two datasets are 90.15% and 86.03% respectively. The performance of CS is better than IM2GPS and SC. We can conclude that both global and local visual features are contributive in image location estimation. Our SLW further gets some improvement over our previous work CS [4]. This shows that the spatial layer information information is worth exploiting for improving the image location estimation performance.

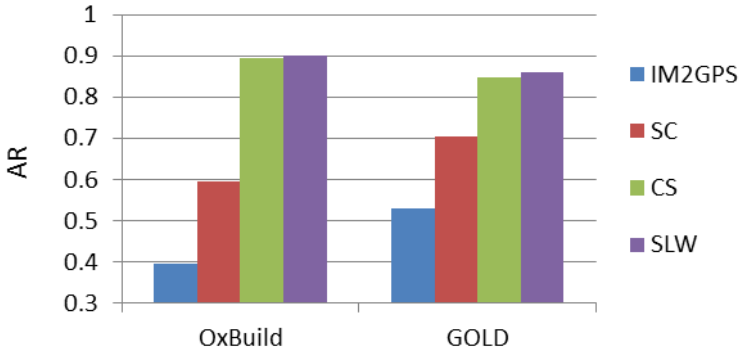


Fig. 5. ARs of IM2GPS, SC, CS and SLW

6.3 Discussion

The performance of our approach is influenced by several main factors. Hereinafter, we discuss their impacts respectively by carrying out a set of experiments.

6.3.1 The Impact of Using Global Features or Local Features

In our experiments, when mining a cluster, the global features are utilized. For each image in a cluster, we gather its R most similar images across the entire training set based on the similarity of the global visual features. We conduct an experiment that in this process, local feature is used instead of global features. We extract local features scale-invariant feature transform (SIFT). A SIFT feature consists of a 128-D

descriptor vector and a 4-dimensional DoG key-point detector vector (x, y, scale, and orientation). We can see from Table 1 that performance improvement is not obvious. The reason is that the clusters are mined for location estimation. We further obtain refined locations of input image via cluster selection. After this process, we just want to generate refined locations of the input image to narrow the scope of retrieval. Moreover, the time cost of local features is certainly more. So, global features are utilized in our work.

Table 1. Average Recognition Rates (%) of using global features and local feature in cluster mining

Dataset	Global features	Local feature
OxBuild	90.15%	90.17%
GOLD	86.03%	86.29%

6.3.2 The Impact of Using Useful Features or All Features

In the part of local feature refinement, salient features selection of images is carried out. For different images, their visual words have different weights during location estimation. The performances of using all features and useful features are discussed here. It can be seen from Table 2 that the performance of using all features is interior to using useful features. So selecting salient words is of significance for image retrieval.

Table 2. Average Recognition Rates (%) of using all features and useful features

Dataset	All features	Useful features
OxBuild	89.77%	90.15%
GOLD	85.51%	86.03%

6.3.3 The Impact of Number of Layers in SLW n

In the part of spatial layer matching, we build spatial layers of visual words for each useful word of those refined images. In our experiments, we build n layers for each visual word. The impact of layer number n to location estimation is discussed here. The AR values of SLW on GOLD are 70.56%, 86.03%, 86.97% and 63.44% respectively when n={1,3,5,7}. It can be seen from Table 3 that with the increase of n the AR is first increasing and then into decline. When n is in the range of [3, 5], better performance can be achieved. If the distance of two visual words is larger than the proper value, their correlation is obviously weaker. During our experiments, n is set to be 3.

Table 3. Average Recognition Rates (%) of different values of n

Dataset	n=1	n=3	n=5	n=7
OxBuild	76.82%	90.15%	91.21%	70.58%
GOLD	70.56%	86.03%	86.97%	63.44%

6.3.4 The Impact of Percentage of Cluster Location Candidates V

In cluster location estimation, for a cluster, we obtain the ranking list of L locations. We select V percent of the L locations as location candidates related to the cluster. Here, we discuss the impact of V to image GPS estimation performances on both GOLD and OxBuild as shown in Figure 6. The AR values of SLW on GOLD is 39.41%, 60.39%, 75.19%, 86.03%, 86.73% and 61.32% with the increase of V . $V$is set 60 in our experiments. If the V is too large, more unrelated locations will be taken into consideration. If the V is too small, the related location will be cut out, which has a worse impact on performance. It can be seen from Figure 6 that with the increase of V, the AR is first increasing and then into decline.

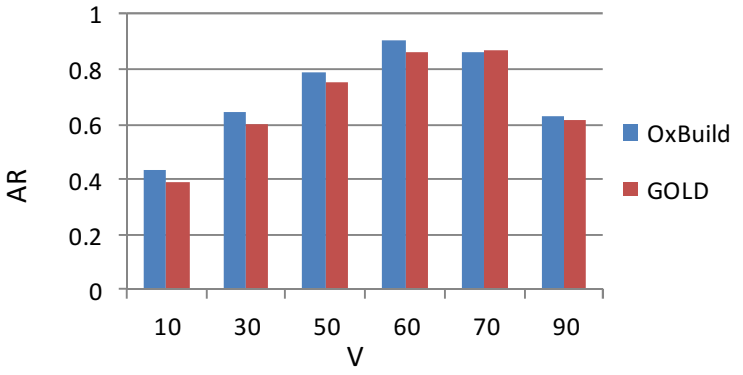


Fig. 6. Impact of cluster location candidates V

7 Conclusion

In this paper, we present a method for image location estimation. In our work, first the images are clustered relying on global features. And each cluster is seen as a single observation so as to mine the relationship among those clusters and locations. Then refined locations are further selected via a cluster selection strategy online. Afterwards, spatial information of visual words is mined. We build spatial layers of visual words (SLW) for further matching, which are generated by involving visual words and the neighboring visual words. The final location estimation is yielded via an online spatial layer matching process. Experiments show that our proposed SLW has better results.

Acknowledgments. This work is supported partly by NSFC No.61173109, No.61128007, No.60903121, Microsoft Research Asia, and Fundamental Research Funds for the Central Universities.

References

[1] Zhang, S., Huang, Q., Hua, G., Jiang, S., Gao, W., Tian, Q.: Building Contextual Visual Vocabulary for Large-scale Image Applications. In: MM 2010, October 25–29 (2010)
[2] Li, Y., Crandall, D.J., Huttenlocher, D.P.: Landmark Classification in Large-scale Image Collections. In: ICCV 2009 (2009)
[3] Zhang, Y., Jia, Z., Chen, T.: Image retrieval with Geometry-Preserving visual phrases. In: CVPR (2011)
[4] Li, J.J., Qian, X.X., Tang, Y.Y., Yang, L.L., Mei, T.T.: GPS estimation for places of interest from social users' uploaded photos. IEEE Trans. Multimedia (2013)
[5] Lowe, D.G.: Distinctive Image Features from ScaleInvariant Keypoints. In: HCV (2004)
[6] Liu, H., Mei, T., Luo, J., Li, H., Li, S.: Finding perfect rendezvous on the go: accurate mobile visual localization and its applications to routing. In: ACM Multimedia (2012)
[7] Gavves, E., Snoek, C., Smeulders, A.: Visual synonyms for landmark image retrieval. In: CVIU (2011)
[8] Zhou, W., Lu, Y., Li, H., Song, Y., Tian, Q.: Spatial Coding for Large Scale Partial-Duplicate Web Image Search. In: MM 2010 (2010)
[9] Hays, J., Efros, A.A.: IM2GPS: estimating geographic information from a single image. In: CVPR (2008)
[10] Chum, O., Philbin, J., Sivic, J., Isard, M., Zisserman, A.: Total recall: automatic query expansion with a generative feature model for object retrieval. In: ICCV (2007)
[11] Wang, J., Yang, J., Yu, K.: Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories. In: CVPR 2006 (2006)
[12] Wu, Z., Ke, Q., Isard, M., Sun, J.: Bundling features for large scale partial-duplicate web image search. In: 2009 IEEE Conference on CVPR, pp. 25–32 (2009)
[13] Chen, J., Feng, B., Zhu, L., Ding, P., Xu, B.: Effective near-duplicate image retrieval with image-specific visual phrase selection. In: ICIP 2010 (2010)
[14] Gavves, E., Snoek, C., Smeulders, A.: Visual synonyms for landmark image retrieval. In: CVIU (2011)
[15] Xue, Y., Qian, X., Zhang, B.: Mobile image retrieval using multi-photo as query. In: ICMEW (2013)
[16] Han, J., Xu, M., Li, X., Guo, L., Liu, T.: Interactive Object-based Image Retrieval and Annotation on iPad. Multimedia Tools and Applications 72, 2275–2297 (2014)
[17] Qian, X., Liu, X., Zheng, C., Du, Y., Hou, X.: Tagging photos using users' vocabularies. Neurocomputing 111, 144–153 (2013)
[18] Liu, X., Qian, X., Lu, D., Hou, X., Wang, L.: Personalized Tag Recommendation for Flickr Users. In: Proc. ICME, pp. 1–6 (2014)
[19] Qian, X., Liu, G., Guo, D., Li, Z., Wang, Z., Wang, H.: Object Categorization using Hierarchical Wavelet Packet Texture Descriptors. In: Proc. ISM 2009, pp. 44–51 (2009)