Scene Location Guide by Image-Based Retrieval

I-Hong Jhuo[1,2], Tsuhan Chen[3], and D.T. Lee[1,2]

[1] Department of Computer Science and Information Engineering,
National Taiwan University, Taipei, Taiwan
[2] Institute of Information Science, Academia Sinica, Taipei, Taiwan
[3] School of Electrical and Computer Engineering,
Cornell University, Ithaca, NY 14853, USA

Abstract. In this paper, we propose a new image-based algorithm to identify where a tourist is when visiting unfamiliar places. When the tourist takes a photo of an unfamiliar place, our algorithm can recognize where the tourist is by retrieving similar images from an image database, where location information is associated with each image. Our method is not only fusing global and local information but using a coarse-to-fine three-stage search process. We first extract image descriptors from the image taken by the tourist and retrieve a number of most relevant images from the database. Then, we re-rank these relevant images based on geometric consistency. Finally, our method determines where the tourist is by using an image-to-class distance measure. Promising performance of the proposed algorithm is demonstrated by the experiments.

Key words: Image Retrieval, Image-to-Class, Information Fusion

1 Introduction

"Where are we?" It is a frequently asked question when tourists visit unfamiliar places. Since camera-equipped mobile devices are now almost ubiquitous, it seems applicable to utilize these devices to identify where tourists are with an image-based positioning system.

With such devices, tourists can take a photo of a prominent building or a scene spot. This image can be used as a query to find similar images from a database consisting of pre-collected images using content-based image retrieval (CBIR) techniques [18, 20, 22]. CBIR techniques analyze an image in its context, where shapes, colors, textures, or any other features may be used, to retrieve similar images from a collection on the basis of syntactical image features. Recently, CBIR methods extract salient features as multi-dimensional descriptors from images and then cluster these descriptors into vocabularies of bag-of-word (BoW) [11]. In fact, such BoW feature has become a dominant representation of images for both object categorization and scene classification [11, 5, 16, 10, 2]. The assumption of bag-of-word is that a different scene can generally maintain the co-occurrence of a number of visual components which play as the role of 'visual textures'. Therefore, an image is composed of a collection of local features which are computed on interest points or on points in a densely sampled grid. The orderless model has a successful application to scene classification and achieves promising performance. However, such models still focus on local features and ignore global information about spatial information. It may limit its descriptive abilities.

There are other retrieval methods for location identification in the literature. A hybrid keyword-and-image system benefiting from both modalities is proposed in [21]. Based on a data-driven scene matching approach, a simple algorithm is proposed to estimate a distribution over geographic locations from a single image [8]. Multi-view geometric feature-based matching approaches have also been applied to location recognition [19, 13, 17]. Specially, the holistic image analysis provides another view to recognize [14, 15]. For scene perception, Oliva and Torralba proposed a "shape of a scene" concept which regards a scene as an object with a unitary shape. They found that the shape of a scene, which is spatial information, can be inferred from its spatial layout and plays an important role in scene understanding. The concept of *spatial envelope* achieved promising performance on the scene classification. In other words, they all provide different viewpoints for visual processing.

In order to consider both the local and global properties, we try to fuse salient features that include not only global spatial information but also local feature properties for improving retrieval abilities in our experiment. Furthermore, we also present a hierarchical approach, which is a coarse-to-fine image-based method for gradually filtering out the irrelevant images in building identification. There are three stages in our method. In the first stage, we compute and organize each image's contextual information for coarse irrelevant image filtering. In the second stage, we apply the RANSAC algorithm [6] to refine correspondence matching and re-rank the relevant images retrieved in the first stage. In the final stage, the location of the query image is determined by an *image-to-class* [1] distance measure, where each distinct location of the relevant images is represented as a distinct class.

This paper is organized as follows: in Section 2, we will describe how to fuse global and local information for roughly filtering out irrelevant images. In Section 3, we will exploit geometric consistency for refining correspondence matching and re-ranking the remaining images. In Section 4, we will apply an *image-to-class* method to determine the location of the query image. And in Section 5, we will show our experimental results and compare our approach to existing ones. Finally, the conclusion will be depicted in Section 6.

2 Stage I: Coarse Irrelevance Filtering

The goal of this stage is to filter out from the database as many images irrelevant to the query as possible. With a restricted computational cost at this coarse stage, we only require a low *false negative* rate of the excluded images since the remaining images will be further verified in the successive stages. The task of the stage can be accomplished by thresholding the similarities to the query based on some off-the-shelf image descriptors. Motivated by the fact that it is generally

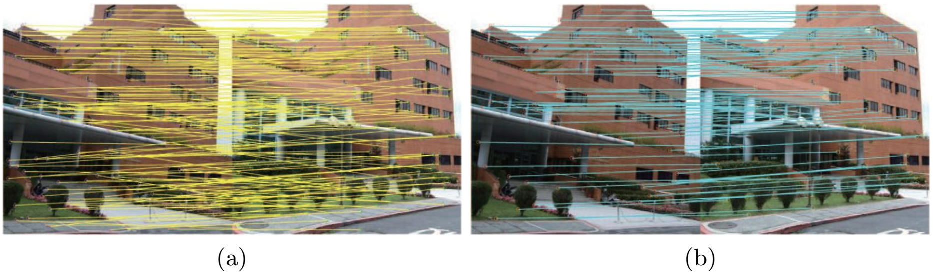

Fig. 1. (a) Two frames showing the same building from different camera view points, with feature points extracted by Harris corner detection, and their correspondences, and (b) final geometrically consistent points as evaluated by the RANSAC algorithm.

difficult to find a descriptor which gives good performances for all images in a large dataset, we adopt several image descriptors for feature extraction, each of which captures distinctive visual cues, such as shape, color, and some perceptually sensitive properties. In the following, we first give a brief introduction to the three adopted descriptors, and then provide an effective way for their combination.

2.1 Pyramid HOG

Shape-based features provide a strong evidence for image categorization, and this phenomenon has been reported in the literature of object recognition. To utilize the discriminant power of shape information, we adopt the *PHOG* (pyramid histogram of oriented gradients) descriptor [3] for shape feature extraction. We set the number of bins of the histograms to 8 and the number of levels in the pyramid to 3, and use χ^2 distance to measure the dissimilarity between two images.

2.2 Color Histogram

Inspired by image retrieval literature, we also consider the use of color information for image similarity measure. It makes sense in the application since many buildings have their distinctive distributions over colors. To this end, we implement color histograms in CIE *Lab* color space, and set the numbers of bins to 21, 40, and 40 for channels L, a, and b respectively. For images under the representation, χ^2 distance is also used as the dissimilarity measure.

2.3 Gist

We adopt the *gist* descriptor [15] as the third feature for its compactness and high performance. The gist descriptor performs Fourier transform analysis to each individual sub-region of an image, and the image is then summarized by a

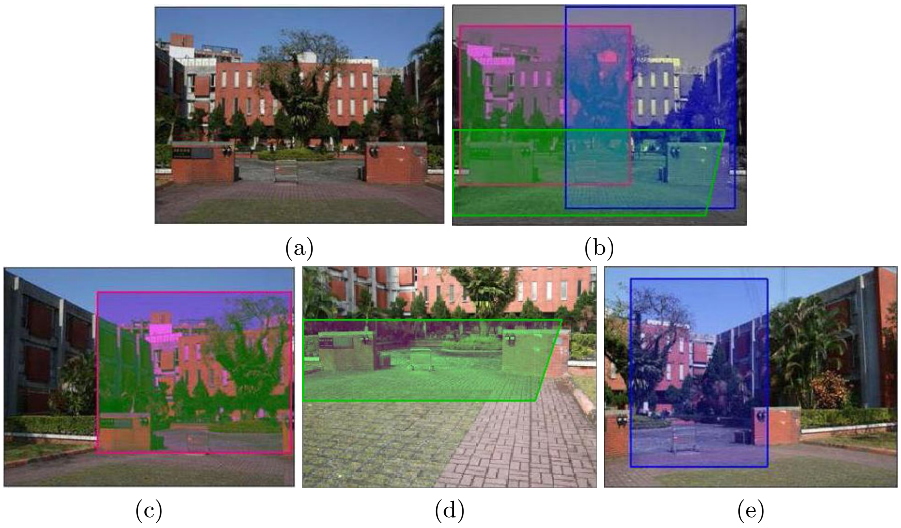

Fig. 2. (a) The query. (b) The responsiveness map of the three images (c $\sim$ e) to the query. It illustrates that none of the three images individually explains the query well, but they do jointly. (c) $\sim$ (e) Three images of the building that pass the first two stages.

set of perceptual properties. The usefulness of gist has been demonstrated in a broad range of applications, such as scene categorization [14] and image retrieval. Euclidean distance is used to estimate the distance between a pair images under gist descriptor.

2.4 Descriptor Fusion

The three image descriptors capture diverse image properties and complement each other. However, the relative importance among them mostly depends on the dataset under consideration. To decide the relative importance for descriptor fusion, we define the distance between the query $\mathbf{q}$ and some images $\mathbf{x}$ in the database as

$$d(\mathbf{q},\mathbf{x}) = d_{PHOG}(\mathbf{q},\mathbf{x}) + \alpha \cdot d_{Lab}(\mathbf{q},\mathbf{x}) + \beta \cdot d_{gist}(\mathbf{q},\mathbf{x}), \qquad (1)$$

where constants α and β determine the weights for the corresponding descriptors. Their optimal values can be decided by using the five-fold cross validation method. By thresholding the distances of images to the query, we can exclude a large portion of irrelevant images while keeping relevant ones.

3 Stage II: Geometric Consistency Checking

Since the distance function in the first stage does not reflect the geometric consistency, we would like to check the geometric consistency by robust feature matching in this stage. We first extract local feature points from the query and the top k relevant images and then perform robust correspondence matching between the query image and each relevant image. We re-rank the top k relevant images according to the number of matched feature points.

Specifically, we detect spatial Harris features [7] from the images. For correspondence matching, we adopt the RANSAC algorithm in [6], where a fundamental matrix is built to remove outliers. The algorithm randomly samples matched points and iteratively estimates the parameters until the fundamental matrix is found.

Figure 1(a) shows the corresponding points without removing outliers, while Figure 1(b) shows corresponding points by the robust matching algorithm. To verify the effectiveness of the re-ranking method in this stage, we use the Normalized Discounted Cumulative Gain (NDCG) [9] measure to evaluate the quality of the top k relevant images before and after the re-ranking. The NDCG score is defined as

$$S_k = \sum_{j=1}^{k} \frac{\delta(c_j, c_q)}{log_2(1+j)} \tag{2}$$

where j is the the index of the jth relevant image, c_j and c_q are the classes of the jth relevant image and the query image, respectively. $\delta(c_j, c_q)$ is 1 if $c_j = c_q$, 0 otherwise.

4 Stage III: Decision Making

Unlike prior stages in which *image-to-image* similarities are computed, the concept of *image-to-class* similarities [1] is employed in the stage. That is, the query is interpreted not by a single image *individually* but instead by the set of survival images *jointly*.

We argue that the computation of image-to-class distances makes sense in our application, and this point is illustrated in Figure 2. The query and the three survivals of the building images, i.e., those that pass the first two stages, are shown in Figure 2a, 2c $\sim$ 2e respectively. According to the *responsiveness map* in Figure 2b, it can be seen that although none of the three survivals individually explain the query well, they do *show* a good match jointly.

To implement the idea, we apply the *DoG* detector [12] to the query $\mathbf{q}$, and use the *SIFT* descriptor [12] to depict each detected point. That is $\mathbf{q} = \{\mathbf{d}_1, \mathbf{d}_2, ..., \mathbf{d}_n\}$, where n is the number of detected points and $\mathbf{d}_i$ is the feature vector of point i. Except for the query, we pre-compute the same representation for each image in the database. Now we are ready to compute the image-to-class distance from query $\mathbf{q}$ to some class (building) c. Let $S = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_m\}$ denote the collection of the feature vectors of detected points in the survival images that belong to class c. Then the image-to-class distance $d_{I2C}(\mathbf{q}, c)$ is defined as follows:

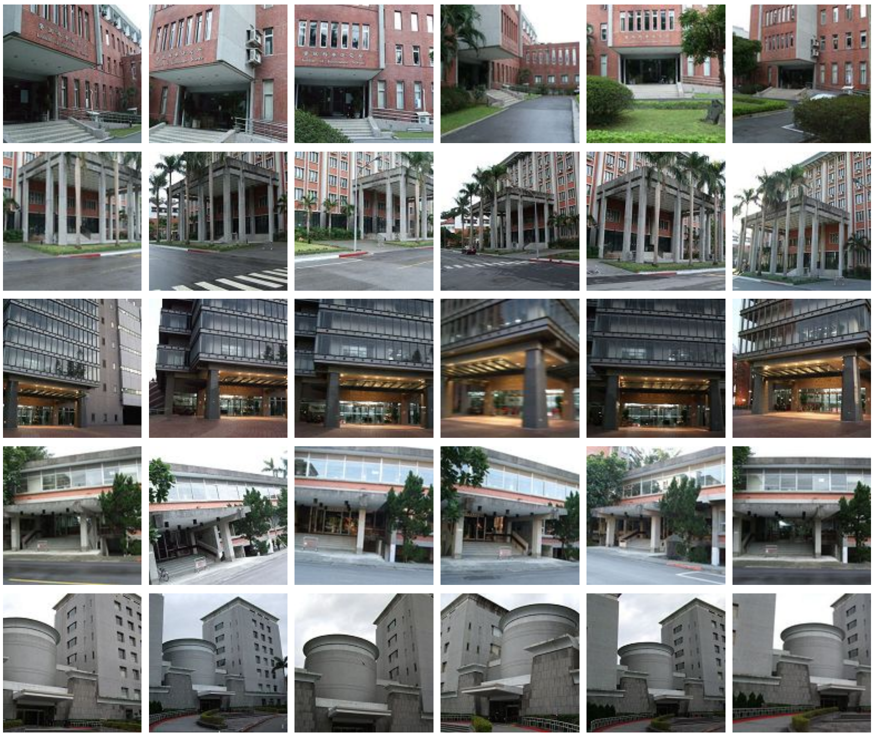

Fig. 3. Sample images in the dataset, each row representing images taken from different views around the same building.

$$d_{I2C}(\mathbf{q}, c) = \sum_{i=1}^{n} \|\mathbf{d}_i - NN_S(\mathbf{d}_i)\|^2, \text{ where} \tag{3}$$

$$NN_S(\mathbf{d}_i) = \arg\min_{\mathbf{x}} \{\|\mathbf{x} - \mathbf{d}_i\|^2 | \mathbf{x} \in S\}. \tag{4}$$

We compute the distances from the query to all the classes in which there are still images left, and complete the prediction by finding the shortest one.

5 Experimental Result

The performance of the proposed approach is evaluated in the section. In the following sections, we describe the used dataset, experiment settings, and quantitative results.

5.1 Experimental dataset

The motivation of our task is to help a tourist identify the location when visiting unfamiliar places. Therefore, we collect an image database that consists

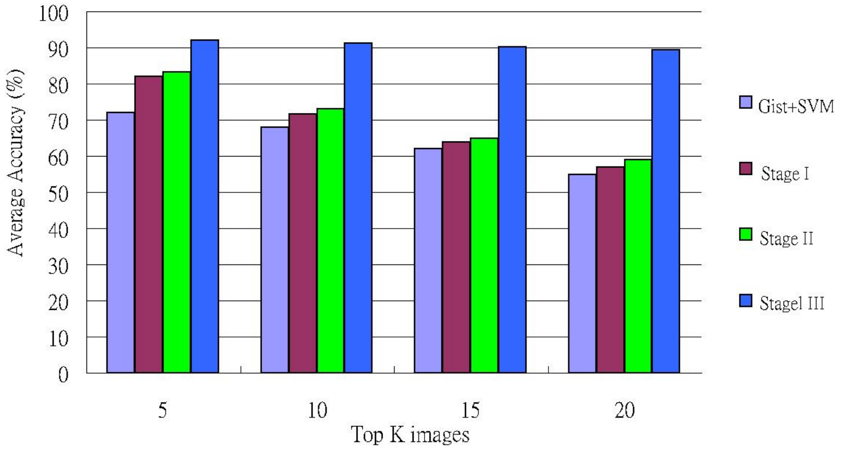

Fig. 4. Accuracy rates broken down into top five, ten, fifteen,and twenty images, respectively. As we can see, image-to-class method present an effective performance at each first k images.

of 13 distinct buildings located in our campus for performance evaluation. The number of images of each building ranges between 75 and 85, and these images have intraclass variations caused by serval different factors, such as photographying viewpoints, scales, and camera settings. We consider images from the same building as relevant, and irrelevant otherwise. The rule serves as the groundtruth. Some examples are given in Figure 3.

5.2 Results

In this section we present experiments for building recognition. We show that our proposed approach gives promising performance over the state of the art method in this task. We concurrently implement global feature representation, Gist with SVM [4] (Gist+SVM) and two kinds of local information, Color with SVM (Color+SVM) and bag-of-features [11] as our baseline in this experiment. In addition, for each image, we build joint histograms of color in CIE L*a*b color space.

To analyze the effectiveness of the three-stage retrieval process, we report the performance stage by stage. All the results in the following are measured by using five-fold cross validation. In the first stage, the query image is compared to those in the dataset, and the most similar k images are selected. Then these k images are re-ranked by confirming their geometric consistency to the query in the second stage. By setting the value of k to 5, 10, 15, and 20 respectively, the average NDCG scores, defined in (1), are reported in Table 1. According to the scores, many irrelevant images are filtered out in the first stage, and geometric consistency checking in the second stage is helpful for score improvement.

Table 1. NDCG Score

Top k	5	10	15	20	25	30
Stage I	2.323	3.241	3.811	4.305	4.845	5.245
Stage II	2.354	3.303	3.883	4.415	4.932	5.313

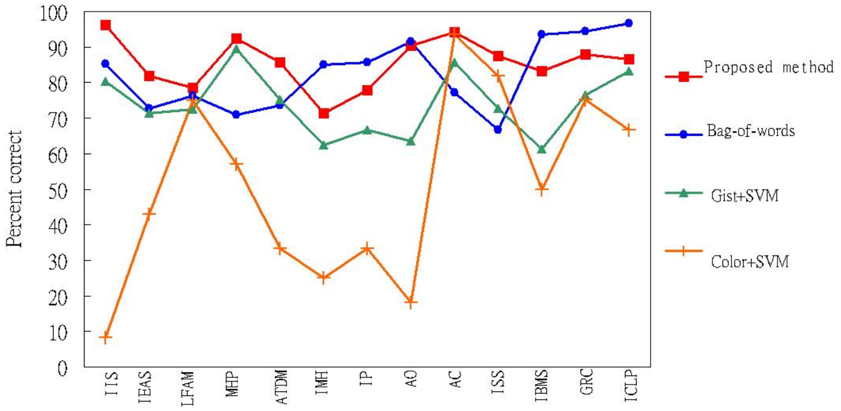

Fig. 5. Comparison with "Gist+SVM", "Color+SVM" and bag-of-words with our dataset. Average performance for the different methods are: our proposed approach: 86.26%, bag-of-words: 82.29%, "Gist+SVM": 73.94%, and "Color+SVM":50.79%

In Figure 4, we report the accuracy rates of both "Gist+SVM" approach and our proposed method in the three stages respectively. Obviously, the usage of image-to-class distances in the third stage significantly improves the effectiveness of our approach. Although the computational cost of image-to-class distances is relatively high, it doesn't limit the applicabilities of our method since image-to-class distances are computed only for images that pass the first two stages, instead of for the images in the whole dataset. We compare our strategy with the state of the art model for this task. For the "Gist+SVM", we use four-fifth images of each class/building for training and one-fifth images for testing. As we can see from Figure 4, our approach fusing local and global information shows promising performance.

The confusion matrix in Table 2 depicts the classification performance of our proposed approach. The average classification performance is 86.26%. We compare our result with those of the state of the art approaches including bag-of-words [11, 5, 16, 10, 2], "Gist+SVM" [14, 15] and "Color+SVM" approaches in Figure 5. The classification results by "Color+SVM", "Gist+SVM" and bag-of-words are 50.79%, 73.94% and 82.29%, respectively. As we can see, our approach achieves better performance over other methods.

Table 2. Confusion Matrix of the proposed approach on the thirteen buildings. The average classification rates of each building are shown in the diagonal and the average accuracy rate of retrieval is 86.26%.

	IIS	IEAS	LFAM	MHP	ATDM	IMH	IP	AO	AC	ISS	IBMS	GRC	ICLP
IIS	96.1	0	10.67	0	0	0	0	3.33	0	0	7.33	6.33	0
IEAS	0	81.81	0	0	0	14.52	15.63	0	0	0	3.47	0	0
LFAM	0	0	83.74	0	0	0	0	0	0	0	2.41	0	0
MHP	0	0	0	92.30	0	0	0	0	3.06	0	0	0	0
ATDM	0	0	0	0	83.71	0	0	0	0	0	0	0	0
IMH	0	1.24	0	0	0	76.3	0	6.19	0	0	0	0	0
IP	0	11.11	0	0	0	0	77.78	0	0	0	0	0	0
AO	0	0	0	0	0	8.05	5.86	90.33	0	0	0	3.18	0
AC	0	0	0	0	0	0	0	0	94.25	0	0	0	0
ISS	0	0	0	0	9.42	0	0	0	0	87.5	2.47	2.54	3.12
IBMS	3.13	5.34	4.31	5.53	0	0	0	0	0	12.11	83.33	0	9.33
GRC	0	0	0	0	0	0	0	0	0	0	0	87.58	0
ICLP	0	0	0	2.07	6.17	0	0	0	2.63	0	0	0	86.67

Figure 6 shows an illustration of our proposed approach. The query image is given in the first column and the first row shows the top 5 images selected by employing both local and global information described in Section2. The second row shows the top 5 images after executing RANSAC algorithm and re-ranking. The third row shows the result via the image-to-class method.

In summary, we have demonstrated the effectiveness by fusing both global and local information for scene building recognition. The decision making in our third stage improves the accuracy and outperforms the state of the art approaches in this task.

6 Conclusion

In this paper, we have presented an image-based approach for recognizing building based on fusion of local and global information. Our coarse-to-fine approach first filters out irrelevant images in first two stages. In decision making stage, image-to-class method is employed based on Euclidean metric for evaluating distances and determines the result of the query image from the top k survival images. We compare our approach with the Color+SVM, bag-of-word approaches and the Gist+SVM method which use local and global information, respectively. The proposed approach shows promising results with respect to the state of the art methods. In the future, we would like to conduct different experiments, such as a larger number of categories and achieve better performance.

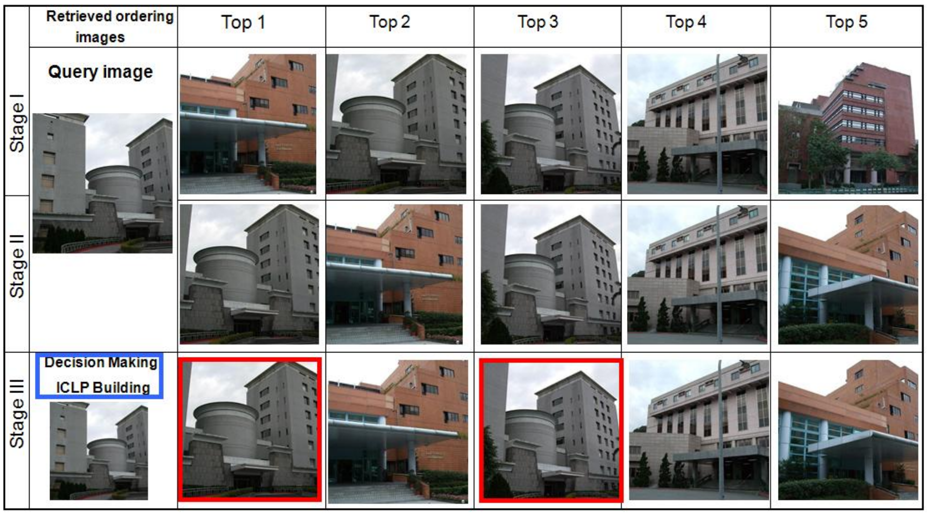

Fig. 6. Illustration of retrieved images by our three stage approach, respectively. The first row shows the top 5 images via stage I. The second row represents the result after re-ranking images of stage I. Final row shows the result that the query image belongs to 'ICLP' building, since the class shown in red-lined box has the shortest distance (d_{I2C}) among the three classes of buildings.

References

1. Boiman, O., Shechtman, E., Irani, M.: In defense of nearest-neighbor based image classification. In: IEEE Computer Society Conference Vision and Pattern Recognition (2008)
2. Bosch, A., Zisserman, A., Munoz, X.: Scene classification via plsa. In: European Conference on Computer Vision. LNCS (2006)
3. Bosch, A., Zisserman, A., Munoz, X.: Image classification using random forests and ferns. In: IEEE International Conference on Computer Vision (2007)
4. Chang, C., Lin, C.: Libsvm: a library for support vector machines (2005)
5. Dance, C., Willamowski, J., Fan, L., Bray, C., Csurka, G.: Visual categorization with bags of keypoints. In: European Conference on Computer Vision International Workshop on Statistical Learning in Computer Vision. LNCS (2004)
6. Fishler, M., Bolles, R.: Random sample consensus: a paradigm for model fitting with application to image analysis and automated cartography. Communications of the ACM 24, 381–395 (June 1981)
7. Harris, C., Stephens, M.: A combined corner and edge detector. In: Proceedings of the Alvey Vision Conference. pp. 147–152 (1988)
8. Hays, J., Efros, A.A.: IM2GPS: estimating geographic information from a single image. In: IEEE Computer Society Conference Vision and Pattern Recognition (2008)
9. Jarvelin, K., Kekalainen, J.: Cumulated gain-based evaluation of ir techniques. ACM Transactions on Information Systems (TOIS) 20, 422–446 (2002)

10. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: IEEE Computer Society Conference Vision and Pattern Recognition. pp. 2169–2178 (2006)
11. Li, F.F., Perona, P.: A bayesian hierarchical model for learning natural scene categories. In: IEEE Computer Society Conference Vision and Pattern Recognition (2005)
12. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision 2, 91–110 (2004)
13. Luo, Z., Li, H., Tang, J., Hong, R., Chua, T.S.: ViewFoucus: Explore places of interests on google maps using photos with view direction filtering. In: ACM, Proceeding of Multimedia. ACM (2009)
14. Oliva, A., Torralba, A.: Modeling the shape of the scene: A holistic representation of the spatial envelope. International Journal of Computer Vision 42, 145–175 (2001)
15. Oliva, A., Torralba, A.: Building the gist of a scene: The role of global image features in recognition. Progress in Brain Research 155, 23–26 (2006)
16. Philbin, J., Chum, O., Isard, M., Sivic, J., Zisserman, A.: Object retrieval with large vocabularies and fast spatial matching. In: IEEE Computer Society Conference Vision and Pattern Recognition (2007)
17. Schaffalitzky, F., Zisserman, A.: Multi-view matching for unordered image sets, or How do I organize my holiday snaps?. In: European Conference on Computer Vision. LNCS (2002)
18. Smeulders, A.W.M., Worring, M., Santini, S., Gupta, A., Jain, R.: Content-based image retrieval at the end of the early years. IEEE Transactions Pattern Analysis and Machine Intelligence 22, 1349–1380 (December 2000)
19. Szeliski, R.: Where am I? : ICCV 2005 computer vision contest. (`http://researchmicrosoftcom/iccv2005/Contest/`)
20. Vogel, J., Schiele, B.: Semantic modeling of natural scenes for content-based image retrieval. International Journal of Computer Vision 72, 133–157 (2007)
21. Yeh, T., Tollmar, K., Darrell, T.: Searching the web with mobile images for location recognition. In: IEEE Computer Society Conference Vision and Pattern Recognition (2004)
22. Zhang, H., Low, C., Smoliar, S., Wu, J.: Video parsing, retrieval and browsing: an integrated and content-based solution. In: ACM, Proceeding of Multimedia. pp. 15–24. ACM (1995)