Accurate Location Identification by Landmark Recognition

Hou Jian, Chua Tat-Seng

School of Computing
National University of Singapore
Singapore
E-mail: (houj, chuats)@comp.nus.edu.sg

ABSTRACT

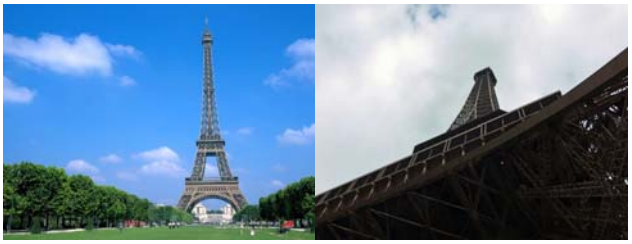
As one of the most interesting scenes, landmarks constitute a large percentage of the vast amount of scene images available on the web. On the other hand, a specific "landmark" usually has some characteristics that distinguish it from surrounding scenes and other landmarks. These two observations make the task of accurately estimating geographic information from a landmark image necessary and feasible. In this paper, we propose a method to identify landmark location by means of landmark recognition in view of significant viewpoint, illumination and temporal variations. We use GPS-based clustering to form groups for different landmarks in the image dataset. The images in each group rather fully express the possible views of the corresponding landmark. We then use a combination of edge and color histogram to match query to database images. Initial experiments with Zubud database and our collected landmark images show that is feasible.

Keywords: landmark recognition, location identification, viewpoint, illumination

1. INTRODUCTION

With the availability of a rapidly increasing number of location-labeled images on the web, it is possible for us to infer geographic information from a given image. This location estimation problem for all kinds of images has been studied in [1] and encouraging results were achieved. For certain scenes like seaside, desert, etc., it is often too difficult or impossible to distinguish between two images taken from different locations. For these kinds of scenes, which are scattered all over the world, the location estimation precision is quite low. In this paper, we will address the problem in a relatively easy but equally important case – limiting the images to the range of landmarks. There are two reasons for doing so. Firstly, people tend to take photos of famous landmarks rather than common scenes and a large percent of images on the web are of this kind. Secondly, landmarks usually have some specific characteristics in appearance that makes them less likely to be similar to surrounding scenes and other landmarks. These two reasons make the location identification based on landmark recognition a necessary and feasible task.

In the landmark recognition problem, one large challenge comes from large variations of viewpoints and illumination. People take photos of one landmark from different viewpoints and in different lighting conditions. If the difference is too large, two images of the same landmark may be seen not to be similar (see Figure 1). These complex conditions have not been fully investigated [2] and pose a lot of difficulties for current recognition algorithms and seem to indicate that one image is usually not enough to fully express a landmark in the database – we need more images of the same landmark. Fortunately, we have hundreds or more of pictures taken from different viewpoints and in different illuminations for most of the landmarks on the web. Generally, these images will cover most of the possible viewpoints and illumination and form a rather full expression of landmarks. Using a set of images rather than one image to express a landmark allows us to alleviate the difficulty in recognition to a large extent.



(a) Example of one landmark in different viewpoints

(b) Example of one landmark in different illumination

Fig.1: Examples of the same landmark in different viewpoints and illumination

The rest of the paper is organized as follows. In Section 2 we describe the method for building the image database and Section 3 presents the details of the image matching algorithm. Experimental results using two databases are shown in Section 4 and Section 5 concludes the paper.

2. BUILDING LANDMARK DATABASE

In this stage we need to set up an image database of a large number of landmarks. The images in the database should belong to landmarks and be GPS-tagged. We can download millions of such images from Flickr.com.

After we have obtained the images, we need to cluster images belonging to different landmarks into different

groups. These groups form a full expression of the corresponding landmark. As mentioned above, images of one landmark from web may be captured in so many different conditions that some of them may not be similar to each other in appearance. Hence the clustering of images based on visual information is usually not reliable for this problem. On the other hand, geographic information in the form of GPS location provides an effective method to distinguish landmarks at different locations. This is the reason why we choose GPS information for clustering.

In some cases, two landmarks may be close to each other (like the bird nest and water cube of Olympic stadium in Beijing) and thus have similar GPS location. Therefore after GPS based clustering we need to further perform a visual-guided clustering in each group to decide if there exist two or more landmarks. For location identification purpose this step is not necessary as GPS location precision is already very high.

There are thousands of landmarks scattered globally and on average there may exist hundreds of images of each landmark. Collecting all the images of one landmark on the web not only create large computation load, but is unnecessary. Generally the variation of viewpoints and illumination is limited and current recognition algorithms have a certain tolerance to these variations. Thus we can divide all the viewpoints into several representative ones, like capturing from left, front, right, bottom, top, etc. In the same way, the illumination is divided into clusters of sunshine, cloudy, rainy, foggy, night with artificial light, etc. Finally we select one image from each sub-group as representative images and they form a new group of smaller size. All these new groups form a new database of smaller size. The process can be depicted in Figure 2.

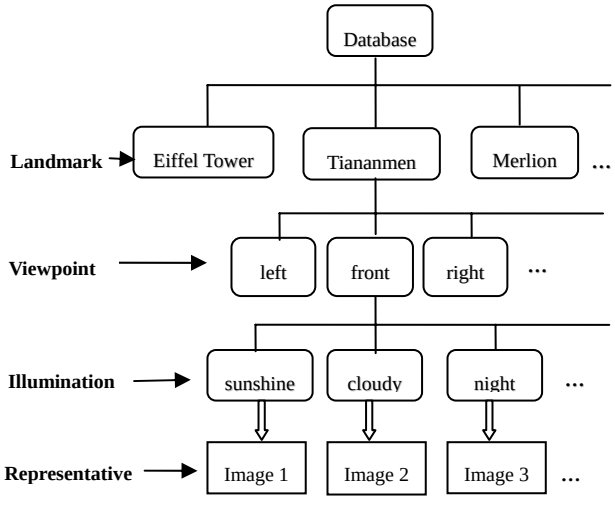


Fig.2: The flow chart of database clustering

With images captured from different viewpoints, we obtain an equivalent 3D model of the landmarks. Moreover, with images captured in different illumination, we get a forth dimension information of the landmarks. These abundant information provides the possibility of accurate landmark recognition in complex conditions.

3. IMAGE MATCHING

3.1 Image Feature

Many features have been proposed in recognition algorithms. Commonly used features include color histogram [3], MPEG-7 EHD [4], gist [5], SIFT [6], shape context [7] etc. Some experimental evaluations reported that SIFT exhibits good performance in the tests on several publicly available image databases [8] [9]. However, as a local descriptor, SIFT based method has the disadvantage of large computation and storage requirement. On the other hand, color histogram achieves comparable performance with SIFT in many of these tests. A variant of color histogram, the so called "localized color histogram" [10], showed better accuracy than SIFT in experiments using Zubud database. The localized color histogram computes the color histogram only on edges whose gradient orientation complies with main vanishing direction, and thus weakly encodes the spatial information. Localized color histogram combines the distribution of edge orientations and colors and achieves large discriminating power. Inspired by this idea, we propose to build a more complete histogram of edge orientation and color distribution. It is composed of three histograms: edge histogram, color histogram and localized color histogram. We briefly introduce them as follows.

Edge histogram is based on the MPEG-7 EHD. We divide the edge gradient orientations into 5 types: horizontal, vertical, 45 degree, 135 degree and non-directional. Separating the image into N×M grids and computing the gradient orientation in each grid, we get a 5-bin edge histogram.

For a color histogram, we adopt the same 1D chromaticity representation as in [10], mainly due to the fact that the hue histogram computed is robust to illumination. The RGB is transformed to (Y, C_b, C_r) in the following form

$$\begin{bmatrix} Y \\ C_b \\ C_r \end{bmatrix} = \begin{bmatrix} 0.2125 & 0.7154 & 0.0721 \\ -0.1150 & -0.3850 & 0.5000 \\ 0.5000 & -0.4540 & -0.0460 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix} \quad (1)$$

Then the hue value is calculated by

$$H = \arctan(C_b, C_r)/\pi \qquad -1 \le H \le 1 \qquad (2)$$

Computing the hue value of each edge grid and quantizing into 16 bins, we get a 16-bin color histogram.

Localized color histogram computes the hue value of the 5 types of edges as in edge histogram. This is different from the method in [10] which uses only the edges whose gradient orientation complies with the main vanishing direction. The reason for doing so is based on the observation that landmarks usually have some special shape and texture and the images tend to include many edges in various directions, unlike common buildings where horizontal and vertical are two dominating edge directions. In this step we obtain an 80-bin localized color histogram.

Before computing the three histograms, we preprocess the images to detect interest zones in the images. Unlike photos taken specially for enjoying the scene, a large number of landmark images are taken by tourists. In these images, landmarks and their surroundings are usually cluttered by people with some poses, other buildings or trees. See Figure 3 for example. Sometimes these "extra" bodies may occupy more space than landmarks. If we include these unrelated information in the histograms, the matching accuracy will be deteriorated.

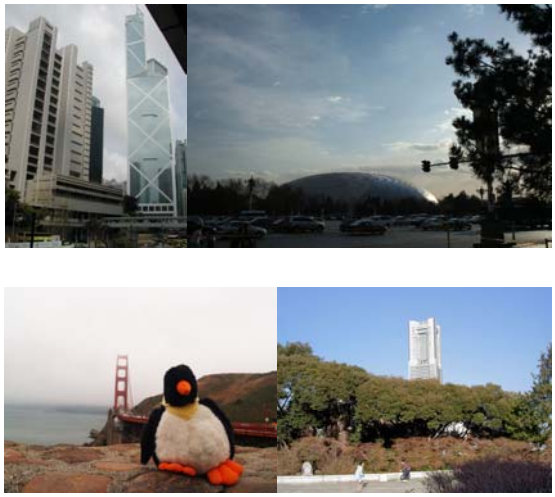


Fig.3: Examples of cluttered surrounding of landmarks

Observation shows that the contour of landmarks can usually be expressed as a skyline. As contour detection is easily affected by cluttered surrounding, we turn to the equivalent features. If we divide the image columns into N bands from left to right and compute the sum of edges, height of edges and range of length in vertical direction of each band, we find the distribution of these measures also show the shape of a skyline (Figure 4). We combine the three measures to get a rectangle region in landmark images as the interest zone.

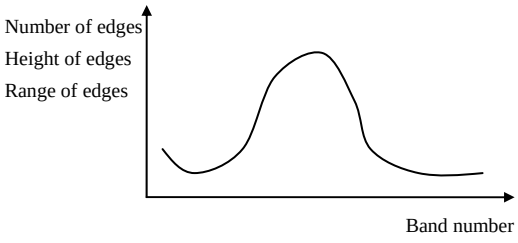


Fig 4: Distribution of three measures in landmark images

In the next step, we first divide the rows in interest zone into 2 bands and compute the localized color histogram in each band, we then divide the rows into 4 bands and compute the edge histogram and color histogram in each band. Thus we get an overall feature histogram of $5\times4+16\times4+80\times2=244$ bins. Here we encode spatial information in the histogram by dividing the interest zones into horizontal bands. We use Figure 5 to show the reason that selecting horizontal bands rather than vertical bands or $N\times M$ grids. In most cases different viewpoints of one landmark are scattered on the ground. That is, images from different viewpoints can be seen as captured by one person moving from left to front and then to right of the landmark. As a result dividing the interest zone into horizontal bands helps to preserve the up-down relation of different edges in the landmark (see Figure 5(a)). However, if we divide the interest zone into vertical bands, a large section appeared in the left band in one viewpoint may appear in the right band in another viewpoint (see Figure 5(b)). The reason for dividing into 4 bands for color and edge histogram while 2 bands for localized color histogram lies in the fact that the latter is a more detailed representation of images and is relatively not so robust to the inaccuracy of interest zone detection.

(a) Interest zone divided into horizontal bands

(b) Interest zone divided into vertical bands

Fig 5: Band dividing in the interest zone

3.2 Similarity measure

After we obtain the feature vectors of images, the next step is to compare the query to the database images. In order to accommodate the possible large difference in texture and color, we compute the similarity of three histograms separately and combine them in the following form

$$S = S_{edge}\cdot w_{edge} + S_{color}\cdot w_{color} + S_{lch}\cdot w_{lch} \quad (3)$$

$$w_{edge} = S_{edge}/(S_{edge}+S_{color}+S_{lch})$$
$$w_{color} = S_{color}/(S_{edge}+S_{color}+S_{lch})$$
$$w_{lch} = S_{lch}/(S_{edge}+S_{color}+S_{lch})$$

where S_{edge}, S_{color} and S_{lch} are the similarity scores, and w_{edge}, w_{color} and w_{lch} are the weights of three histograms in final similarity measure respectively. In this way, if only one feature in color and texture has no large variation between the two images, there will be at least one of the three measures that has a large value and are attributed a large weight. This helps to reduce the effect of large variation of viewpoint and illumination to a certain extent.

For one landmark in the database, each image in its group will give one similarity measure with the query. In our work, the ranking is for groups but not single images. That is, the top N candidates are N different groups. Therefore we need to compute the similarity of each group based on

the similarity scores of all images in the group. In the current stage we select the largest similarity in the group as the measure of the group and use this measure to rank the landmarks. In fact, as images in the same group are more or less similar to each other, it is possible to achieve a better performance by using the overall matching result to rank the groups in a way. We will investigate the method further in later work.

4. EXPERIMENTS

We use two image databases to evaluate the performance of our method. The first is the publicly available Zubud building image database, where the database is composed of 1,005 images of 201 buildings, and the query is 115 images of some of the 201 buildings. Though large variation of viewpoint, illumination and scale is rare, the database provides a fairly large number of buildings of different styles and thus can be used as a simple landmark image database. The other database is composed of our collected landmark images. In the current stage, it is composed of 7,053 images of 55 landmarks from all over the world. The collected images experience most of the possible complex conditions in reality you can imagine: the cluttered surroundings, large variation of viewpoints, illumination and scale, etc. Though still small in size, we see it as a good starting towards a practical system.

4.1 Zubud Database

In Zubud database, some images are captured with camera rotated 90 degree, that is, in these images the sky-ground relation is left-right but not up-down. As the rotation will change the edge gradient direction and thus change the histograms, we pre-rotated these images into the usual orientation. We argue that this kind of images occupy only a small fraction of the images available on the web and are possible to be detected in some way.

Using the proposed method, we archive 95.6% top 1 recognition rate and 97.3% for the top 5 list. This is better than using the localized histogram (90.4% and 96.5% respectively) or SIFT (90.4% and 94.8% respectively) alone, and just slightly lower than the combined localized color histogram and SIFT method (96.5% recognition rate). The result shows that by encoding more spatial information and the compensation of different feature histograms, the proposed method acquires larger discriminating power.

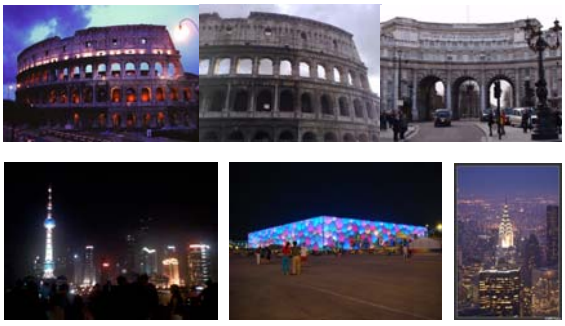
4.2 Collected Landmark Image Database

The images in this database are collected from Flickr.com. It exhibits cluttered surroundings and significant variation of viewpoint, illumination and scale. Figure 6 shows samples of the images of the Big Ben in London. These complex surrounding conditions present challenges to current recognition algorithms in both the feature detection and similarity measure steps. As an initial test, we manually selected four images from each landmark group and form a query of 220 images. These four images are specially selected to cover the extreme viewpoint, illumination and scale conditions. See Figure 7 for samples of the query images. In the tests, the recognition rate for top 1 image is 45% and the top 5 is 85%. Some recognition results are shown in Figure 8. In each examples the first image is the query and the other 5 are the top 5 recognition image. As in the general cases, the images have a small chance to contain the extreme conditions, the performance in practical use should be better than in the tests.

Fig. 6: Samples of images of the Big Ben

Fig.7: Samples of query images

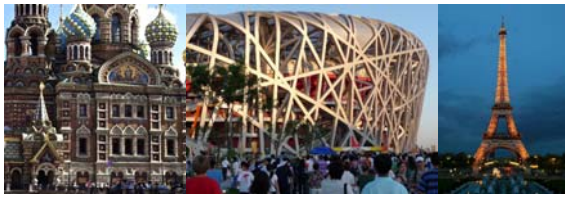
(a)



(b)

(c)

(d)



(e)

(f)

Fig.8: Samples of recognition results

5. CONCLUSTION AND DISCUSSION

If the content of images is totally unconstrained, accurately estimating the geographic information from images is a very difficult and sometimes impossible task. However, if we limit the images to the set of landmarks, then the problem becomes more feasible but is of equal importance. Due to the complexity, the problem of landmark recognition in practical conditions of cluttered surrounding, large variation of viewpoints, illumination and scale has not been fully investigated. In this paper, we proposed a possible solution to the problem and reported some primary work that we have completed.

In order to cope with the large variation of viewpoints and illumination, we proposed to use a group of images captured from different viewpoint and in different illumination to express a landmark. We then selected representative images from each group to form new image groups of smaller size. For the problem of cluttered surrounding and large variation of scale, we proposed to detect interest zone which contains most landmark information with minimum unrelated information. With regard to image features, we showed that by proper use of the spatial distribution of features, combined edge and color histogram produce comparable performance to SIFT based feature.

In the next step, we plan to work in the following directions to complete and refine the system:

(1) Building up landmark image database

The tasks here include the detection of landmark images from GPS-tagged images and the selection of representative images. After using GPS location to cluster images into different groups, we need to detect landmark images and remove the other building and non-building images. Also, we need to select representative images from these groups in order to reduce the database size.

(2) Image feature detection

In order to reduce the harmful effect of the surroundings, it is necessary to detect the interest zone and detect features only in this zone. Since local and global features have their advantages and disadvantages, we plan to combine them to better tackle the peculiarity of landmarks.

(3) Ranking of Landmark groups

Given a query, each database image has a similarity score with the query. As our ranking is based on groups but not single images, the problem is then how to combine the similarity scores of images in one group so as to decide the score of the group. We used the largest similarity in one group to represent the group in current implementation, and will try other combination methods.

(4) Location identification

After we obtain the ranking of each landmark, we must decide the location of the query based on the location of landmarks in the database. Previous work [1] represents the estimated location as a probability distribution. We plan to explore better and more accurate method.

6. REFERENCES

[1] J. Hays, A. Efros., "IM2GPS: Estimating Geographic Information from A Single Image," CVPR '08 pp. 1-8, 2008.

[2] G. Nguyen and H. Andersen et al., "Urban Building Recognition during Significant Temporal Variations," WACV '08 pp. 1-6, 2008.

[3] A. Smeulders and M. Worring et al., "Content-Based Image Retrieval at the End of the Early Years," IEEE PAMI Vol. 22. No.12. pp. 1349-1380, 2000.

[4] S. Chee and K. Dong et al., "Efficient Use of MPEG-7 Edge Histogram Descriptor," ETRI Journal Vol.24. No.1. pp. 23-30, 2002.

[5] J. Hays, A. Efros., "Scene Completion using millions of photographs," SIGGRAPH '07 Vol.26. No.3. pp. 1-7, 2007.

[6] D. Loew., "Distinctive Image Features from Scale-Invariant Keypoints," IJCV Vol.60. No.2. pp. 91-110, 2004.

[7] S. Belongie and J. Malik et al., "Shape Matching and Object Recognition Using Shape Contexts," IEEE PAMI Vol.24. No.4. pp. 509-522, 2002.

[8] K. Mikolajczyk, C. Schmid., "A Performance Evaluation of Local Descriptors," IEEE PAMI Vol.27. No.10. pp. 1615-1630, 2005.

[9] T. Deselaers and D. Keysers et al., "Features for Image Retrieval: An Experimental Comparison," Information Retrieval Vol.11. No.2. pp. 77-107, 2007.

[10] W. Zhang, J. Kosecka., "Hierarchical Building Recognition," Image and Vision Computing, Vol.25. No.5. pp. 704-716, 2007.