INQUIRING OF THE SIGHTS FROM THE WEB VIA CAMERA MOBILES

Yinghua Zhou[1*], Xin Fan[1*], Xing Xie[2], Yuchang Gong[1], Wei-Ying Ma[2]

[1]Department of {CS, EEIS}, University of Sci. & Tech. of China, Hefei, Anhui, 230027, P.R. China
{yhzhou, van}@mail.ustc.edu.cn, ycgong@ustc.edu.cn
[2]Microsoft Research Asia, 5F, Sigma Center, No. 49, Zhichun Road, Beijing, 100080, P.R. China
{xingx, wyma}@microsoft.com

ABSTRACT

In this paper, we presented an image search service for mobile users. It can be used to acquire related information by taking and sending pictures to the server, for example, getting book reviews by a photo of the cover. The key problem here is to find images that contain the same prominent object as that in the query image. In the literature, local feature based image matching has been proven to outperform those based on global features. When using local features, however, one query image may contain thousands of high dimensional feature vectors. Each feature vector needs to match against millions of features in the database. Therefore, it is critical to design an efficient search scheme. Our proposed matching approach was based on identifying semi-local visual parts from multiple query images. Experiments on two real-world datasets showed that this approach was superior to conventional solutions.

1. INTRODUCTION

Mobile phones with embedded cameras are very popular nowadays and have huge growth potentials. Mobile cameras are becoming a promising HCI manner for mobile devices, just like the emergence of mouse for desktop computers. Most current services for information acquisition on mobiles, such as Google Mobile, Google SMS and Yahoo! Mobile, are using text-based inputs. Nevertheless, sometimes it is difficult for users to describe their information requests in words. Instead of current flat query modes, camera phones can support much richer queries, not only text but also images.

In this paper, a web service is designed to support users to inquire the relevant information from the Web via photos of what they see, for example a picture of a book or a hotel. We first cache some Web sites on several specific topics. All the pages and images are stored in advance. If one of the stored Web image can be matched with the query images from camera phones, the information from the corresponding Web pages will be returned to users.

The key problem is to seek for the images that contain the same prominent object or scene as query images, for example, photos of the same book or building captured by different persons or at different time. There are already some researches [2][5][9] on searching in a database with captured pictures using camera phones. Most of previous work uses Content Based Image Retrieval (CBIR) method, which is a coarse-grain matching schema and returns images globally similar to the query image. Local feature based schemes outperform conventional global measurements in robustness and accuracy [1]. However, one single picture may generate hundreds to thousands of features, all of which have to be used as queries to match millions of high dimensional features in a large-scale database. Additionally, the distance computation is time-consuming for these high-dimensional features. Consequently, an efficient search schema is critical for a large-scale image database.

We propose an image matching approach based on identifying semi-local visual parts which are grouped from local features. Considering the fact that the captured photos often contain cluttered background and other unwanted objects, we generate semi-local visual parts from an image sequence or multiple shots to remove these noises. This solution requires users to take more than two pictures or a shot video clip, which is also convenient for users. This approach can significantly reduce the computations for feature vector matching and improve the accuracy of retrieval. The experiment results on different datasets showed that our approach performed better than conventional approach in retrieval performance.

This paper is organized as follows. In Section 2, we introduce the flow chart of our web service. In Section 3, the image matching method based on semi-local visual parts is described in detail. Experiments in Section 4 show the efficiency of our solution. At last, conclusions are given in Section 5.

2. SYSTEM FRAMEWORK

The flow chart of our web service is shown as Figure1, which has two parts: offline process and online process.

[*] This work is performed at Microsoft Research Asia.

In offline process, we first retrieve some Web sites on several specific topics and build a Web image database. Then local features of each image are extracted and indexed. The invariant local descriptors are used extensively to recognize different views of the same object as well as different instances of the same object class. It is empirically found in [8] that the SIFT descriptor [7] shows very good performance, invariant to image rotation, scale, intensity change, and to moderate affine transformations. Though some other affine invariant detectors and descriptors can provide more stable results under larger viewpoint changes, considering the computational cost, we choose the SIFT based method. Features are represented as a vector in a 128-dimensional space.

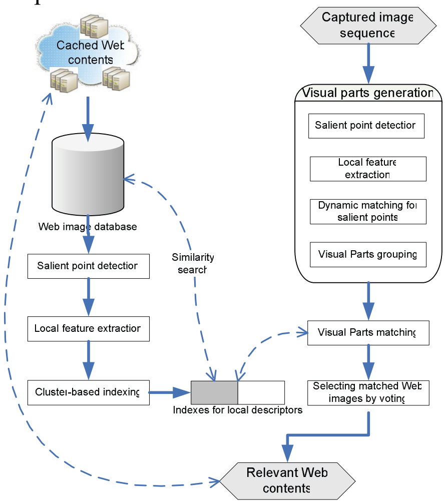


Figure 1. Flowchart of our web service

Index techniques are necessary to speed up the search in a large-scale database. Although there is much work on high dimensional index, almost all methods suffer from the "curse of dimensionality" in high dimensional space. The superiority of cluster-based method over most existing techniques is shown in [3]. In our system, the index structures are based on unsupervised clustering with Growing Cell Structures (GCS) artificial neural network [4], which shows good performance for higher dimensional space.

In online process, an image sequence is captured, and then semi-local visual parts are generated from them before searching in the database and finding images similar to query images. The images with the same prominent objects or scenes are returned by the voting results of matched visual parts. Finally, the web pages in which these images are located can be returned for reference. The detailed process will be described in Section 3.

In this system, two factors are vital for the retrieval performance. One is the number of query features; the other is the index of feature vectors in database. For the former, we propose the method based on semi-local visual parts. This method can effectively and extensively reduce the features from the cluttered background and noise objects. For the index, GCS is used to improve the performance.

As shown in Figure 2, the query procedure can be described in an exemplary scenario. When a mobile user is attracted by a book or a building, and wants more information about it, he can take photos of it and then send the photos to our system by MMS (Multimedia Messaging Service) or email. The user would soon receive relevant Web pages sent out by our service system.

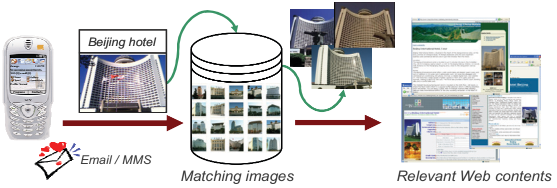


Figure 2. Illustration of an example query procedure

3. VISUAL PARTS

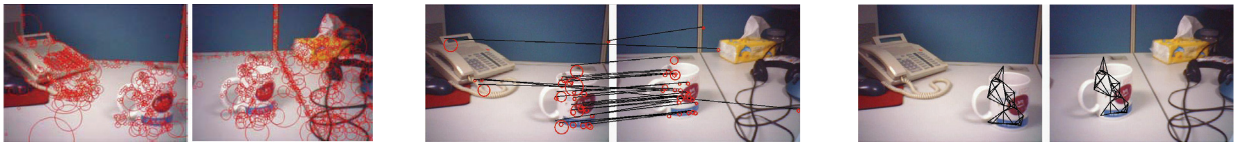
The local configurations of salient points are often stable across different version of images for an identical object [6]. Therefore, the spatial configurations with similar appearance are potential visual correspondences. Inspired by this, we first get the correspondences from multiple query images, and then group the stable neighboring salient point to compose visual parts. Among the local feature matches in the captured photos, a large proportion of them are false matches caused by features from background clutter and other objects. Consequently we often need to identify the correct matches among over 90% outliers [7]. By introducing the visual parts instead of the direct point match, the precision of the retrieval results is considerably improved. In addition, the amount of query points is also remarkably reduced, which saves much computational cost in the process of similar search in the image database.

3.1. Visual parts generation

For the input query, the user simply needs to capture multiple pictures continuously. Then, using these pictures, we extract visual parts by adding the correspondent salient points between each two neighboring input images.

There are mainly three steps as shown in Figure 3. Firstly, SIFT algorithm is used to extract local features from each query image. Each feature is represented by a 128-dimensional vector.

Secondly, we identify the matched features between neighboring query image pairs respectively. The matching of features corresponds to the nearest neighbor (NN) relationship in the 128-dimensional vector space. Here we use the L2 distance to measure NN relationship. But the nearest feature vector is not always the correct correspondence in the large amount of feature vectors. Therefore we choose the two nearest neighbors as candidate matches. The matches would be rejected when the ratio of



(a) Extract local features of query images (b). Match salient points based on local features (c).Generate visual parts by grouping NN salient points

Figure 3. The generation of semi-local visual parts

distance from the closest neighbor to the distance of the second closest is greater than 0.8. In our experiments, it is proven that this method can eliminates 90% of false matches while discarding less than 5% of correct matches.

Thirdly, we select triples of matched salient points which are neighbors in position from one image. Assuming we build up *N* correspondent salient point pairs, for any point *A* of these *N* pairs which belongs to the first image, we can define its local neighborhood region as a circle whose center is this point and radius is a constant factor of its scale (such as 3 in our experiments). Any two points *B, C* of matched pairs whose position locate in the local neighborhood region within the first image are grouped a visual part with *A*. Thus, a visual part can be regarded as a triple of neighboring salient points. All the triples in the first images are enumerated in this neighborhood as the visual parts of the adjacent image pair. This method can effectively suppress the noise salient points since they are often with a smaller scale and no other salient points nearby.

For the multiple query images, we add the visual parts extracted from each image pair to form a visual part set. In this case, more query images require, longer computation time. Considering the tradeoff between the computation time and the performance, we use two query images in our experiments of Section 4.

3.2. Visual parts matching in database

The generated visual parts will be used as inputs to find matched counterparts from the database. There are two steps.

The first step is to get K-NN of each point of query visual parts from the database. Then the returned points are grouped if they belong to the same image in the database. For one database image, the more points are matched, the higher probability the image is similar to the query images. Therefore, we choose those database images of large number of matched points as candidate images. This procedure is similar to the voting mechanism. Thus, according to the matched relationship between salient points, it is easy to get matched parts in each candidate database image.

The time to retrieve in the database is proportional to the number of input query points. The noise points from clutter background are remarkably reduced, so the number of points of query visual parts is much smaller than that of any one query image. The retrieval time is also greatly reduced.

The second step is to validate spatial constraints between these matched triples from the candidate images and the query parts. The triples are rejected if they don't meet affine transform of query parts. The voting is executed based on the number of matched parts in each candidate image in database. The validation based on visual parts is one kind of semi-local spatial constraints which allows for greater range of global shape deformations and local deformations.

4. EXPERIMENTS

We will discuss and analyze the performance of our solution and compare it with conventional algorithms. In conventional image matching algorithms, only one image is used as query image. Then all the extracted feature vectors are input into the database to search NN as matched features. Finally results are returned based on the number of features matched between the query image and each database image. Additionally, the transformation verification based on RANdom SAmple Consensus (RANSAC) is used to validate spatial constraints of these matched points, i.e. to verify the correctness of the matched images.

4.1. Datasets and Settings

To evaluate our solution, we built two datasets, one was book covers that were crawled from amazon.com; and the other was famous buildings in Beijing. For there is no enough database images on the Web, these datasets were captured and labeled by several volunteers. The related data can be downloaded from http://research.microsoft.com/~xingx/ICME06_datasets.html. Query images in our experiments are downscaled to 320x240. The statistics of our datasets and query sets are shown as Table 1.

Table 1. Statistics of our datasets and query sets

	Number of db images	**Number of local features in db**	**Number of query pairs**
Buildings	1259	1.6M	25
Book cover	2557	0.6M	18

4.2. Query time analysis

The computational cost of our method mainly consists of two parts: 1) the generation of visual parts from query images 2) searching matched parts in the database. Most of the time in conventional method is spent on searching in the database for the matched ones of each salient point in the query image.

As shown in Table 2, our method showed better query performance than conventional method. This is because the time to search in the database is proportional to the number of input query points. We greatly reduced the number of query points by generating visual parts from query images.

As shown in Table 3, the average number of query points is reduced from 1,187 to 262 on the building dataset, and from 546 to 171 on book cover datasets. Though we had to spend some time to generate visual parts, our experiments show the time is much smaller than searching all points in a large-scale database.

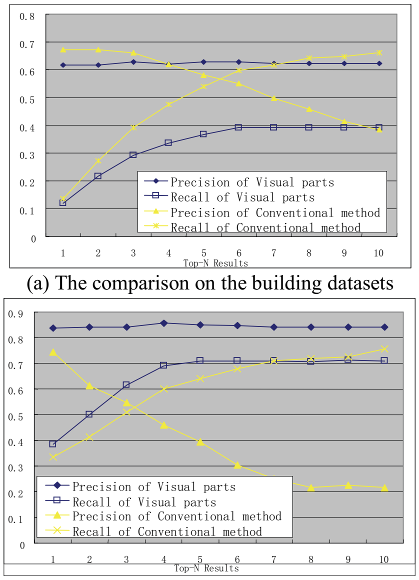
Table 2. Average query time statistics and comparison

Datasets	Visual parts (ms)			Conventional method (ms)
	Parts extraction	Query in db	Avg. query time	Average query time
Building	487.4	875.8	1,363.2	3695.9
Book cover	143.1	350.7	493.8	1108.1

Table 3. The comparison of the average number of query salient points

Datasets	Visual parts		Conventional method
	Avg. number of query parts	Avg. number of query points	Avg. number of query points
Building	1848	262	1187
Book cover	1428	171	546

4.3. Precision and recall analysis



(a) The comparison on the building datasets

(b) The comparison on the book datasets

Figure 4. The precision and recall comparison on different datasets

When the same dominant object appears in both the query image and the result image, we define it as a correct match.

Considering that the screen of mobile device is limited and the content of many web pages about the same object is always very similar, we should first try to maximize the precision and at the same time to guarantee the acceptable recall.

Assuming that the number of returned images to mobile user is no more than Top-N, we compared the performance with the different value of Top-N from 1 to 10. As shown in Figure 4, our solution improved the precision of retrieval results comparing with conventional approaches and at the same time the recall was acceptable. There are two main reasons. One is that visual parts can reduce the number of salient points from background clutter and other objects. The other is that the verification methods are different. The verification of conventional method is based on the RANSAC. Though RANSAC is mostly insensitive to outliers, it will fail if the fraction of outliers is too great. In our solution the verification is based on appearance and geometric matching of parts. It is one kind of semi-local spatial constraints, and is robust to both rigid and non-rigid deformations. Thus, our solution had better performance.

5. CONCLUSIONS

We provided a web service for users to acquire related information via camera photos, such as inquiring book review by sending photos of book cover. The key challenge is to improve the performance of image matching in a large-scale database. We proposed an image matching approach based on semi-local visual parts. The experiments on different datasets showed that our approach can effectively and extensively separate the features from the cluttered background and noise objects. Therefore, it reduced the computations of salient point matching and improved the retrieval performance at the same time.

6. REFERENCES

[1]. Baumberg A., “Reliable Feature Matching across Widely Separated Views”. *In Proc. CVPR*, 2000, pp774-781.

[2]. Fan X., Xie X., Li Z., Li M. and Ma W.Y., “Photo-to-Search: Using Multimodal Queries to Search the Web from Mobile Devices”. *In MIR*, 2005.

[3]. Ferhatosmanoglu H., Tuncel, E. Agrawal D., and Abbadi A. E., “Approximate nearest neighbor searching in multimedia databases”, *In Proc. ICDE*, 2001, pp.503-511, 2001.

[4]. Fritzke B. “Growing cell structures - a self-organizing network for unsupervised and supervised learning”. *Neural Networks*, vol.7 no.9, pp1441-1460, 1994

[5]. Hare J.S. and Lewis P.H., “Content-based image retrieval using a mobile device as a novel interface”. *In Proc. SPIE*, 2005, pp. 64-75.

[6]. Lazebnik S., Schmid C. and Ponce J., “Semi-local Affine Parts for Object Recognition”. *British Machine Vision Conference*, 2004, pp.779-788.

[7]. Lowe D.G. “Distinctive Image Features from Scale-Invariant Keypoints”. *International Journal of Computer Vision*, vol. 60, no.2, pp91-110, 2004

[8]. Mikolajczyk K. and Schmid C. “A performance evaluation of local descriptors”. *IEEE Trans. on Pattern Analysis & Machine Intelligence*, vol. 27, no.10, pp.1615-1630, 2005.

[9]. Yeh T., Tollmar K., Grauman K. and Darrell T., “A picture is worth a thousand keywords: image-based object search on a mobile platform”. *In Proc. CHI*, 2005, pp2025-2028.