

SnapToTell: A Singapore Image Test Bed for Ubiquitous Information Access from Camera

Jean-Pierre Chevallet², Joo-Hwee Lim¹, and Ramnath Vasudha¹

¹ Institute for Infocomm Research (I2R),
21 Heng Mui Keng Terrace, Singapore 119613

² IPAL-CNRS I2R-NUS joined Laboratory,
21 Heng Mui Keng Terrace, Singapore 119613
{joohee, viscjp}@i2r.a-star.edu.sg

Abstract. With the proliferation of camera phones, many novel applications and services are emerging. In this paper, we present the SnapToTell system, which provides information directory service to tourists, based on pictures taken by the camera phones and location information. We present also experimental results on scene recognition based on a realistic data set of scenes and locations in Singapore which form a new original application oriented image test bed freely available.

1 Introduction

Imagine you are at a tourist spot looking at a beautiful monument and instead of searching through your travel guide books to learn more about the scene, you snap a picture of the scene using your camera phone. You phone send it to a service provider via Multimedia Messaging Service (MMS), and little time later you receive an audio clip (MMS) and/or a text message (SMS) that provides you more information about the scene. You can continue to enjoy the scene while your fingers carry out this information retrieval task. We promote this kind of picture-driven information access scenario, known as *SnapToTell* in this poster. As the saying goes “A picture is worth thousand words”, you can forget about the hassle of looking up scene description in document that distracts you from enjoying the scene to access a text-driven information directory. We feel that camera mobile phones will become pervasive personal devices beyond the role of traditional voice communication.

From the technological point of view, obtaining location-based information is already possible with the GPS devices or the GSM cellular network infrastructure¹. However, knowing the location of a mobile phone user is not sufficient to determine what he or she is interested in (or looking at). The location-based information certainly helps to refine the user’s context, but fails to capture his or her intention. The image retrieval aspect is still complementary to the context localization information.

¹ We use only GSM cell id in our current prototype.

2 SnapToTell Paradigm

The application scenario proposed here is similar to the one in [1] : the client device used is a PDA system connected to internet through WLAN. It supposes that this wireless access point is installed in the area in which the system is going to work. The system includes an iPAQ 3870, a NexiCam PDA camera, an orientation sensor, and a GPS receiver. The position detection is ensured by a GPS attached to the PDA. However, the direction and tilt sensor is connected to the PDA via a laptop computer due to technical difficulty. In our case, we have chosen a camera mobile phone (Nokia 7650) which is a lighter and more ubiquitous device. The camera is integrated into the communication device and localization is provided by the telecommunication operator.

In the PDA prototype [1], the image taken from the connected camera together with GPS and orientation data are sent to a server. The server then runs the 3DMax program to generate a reference image from the same position and angle in a 3D model built In SnapToTell, our approach of scene recognition is different. Instead of unnatural matching between a real image and a synthesized image from 3D model, our server will match the query image with *different images of a scene*, taken using different angles and positions. We think that 3D model construction is costly and not applicable to all kinds of scenes.

In essence, our scene indexing and matching strategy assumes that *a set of images of the same object or scene* have in common some characteristic recurrent local features that are discriminative enough to correctly detect the object among other possible ones in a given area. The location information about where the picture was taken reduces the search space and tends to simplify the problem as can be seen in our experiments.

3 Scene Database

Using Singapore as a test, we have set up an original data set of image and descriptions. We have divided the map into 6 zones, 15 locations and 88 scenes. A zone includes several locations, each of which may contain a number of scenes. A scene is characterized by images taken from different viewpoints (average of 17 images per scene), distances, and possibly lighting conditions. Besides a location ID and image examples, a scene is associated with a text description, an audio description which is send to the user an answer to his query.

Using only 530 images in our base ², we have conducted an empirical study on scene recognition. We have adopted color histograms [2] to characterize and index the images. They are known to be invariant to translation and rotation about the viewing axis and change only slowly under change of angle of view, change of scale, and occlusion.

We have experimented with both global and local color histograms, i.e. using image blocs. Spatial locations are sometimes important for discriminating more

² The current state of the base includes 1600 images from 7 camera.

localized objects. Then local histograms will provide good sensitivity to spatial specificity. Furthermore, we can attach different weights to the blocks to emphasize the focus of attention: in our case we have emphasized the center. That is, the similarity λ between a query q (with m local blocks Z_j) and an image x (with m local blocks X_j) is defined as:

$$\lambda(q, x) = \frac{\sum_j \omega_j \cdot \lambda(Z_j, X_j)}{\sum_k \omega_k}, \quad (1)$$

where ω_j are weights, and $\lambda(Z_j, X_j)$ is the similarity between two image blocks defined as

$$\lambda(Z_j, X_j) = 1 - \frac{1}{2} \sum_i |H_i(Z_j) - H_i(X_j)|. \quad (2)$$

We obtain 71% for precision of scene recognition, using 11 bins and 3x3 blocs, and 82% using location, which is enough in practice.

4 Conclusion

In this poster, we present an experimental but fully functional system for mobile and ubiquitous multimedia information retrieval. The histogram is computed in the phone itself, reducing the amount of data to be transferred. As the phone we use does not have floating point, only the raw histogram computation is sent, final normalization is computed on the server. Our approach deals with *real situation* and *real access device* in order to measure the feasibility of such a system. It turns out that we have stretched the limit of currently available wearable technology, but we are convinced that ubiquitous computing is going to have rapid development in the very near future: that will solve some of the technical limitations uncounted.

Results obtained shows that simple matching, based on color histograms, combined with localization information, seems powerful enough to solve this particular image matching problem mainly because of task we have: retrieving among a set of images describing one object, the one that is closed to user one. It is not a usual IR querying task, and the poor value of the precision at full recall is not that significant in this case. The complete test collection is freely available under <http://ipal.imag.fr/SnapToTell.html>.

References

1. W. Mai, G. Dodds, and C. Tweed. Mobile and ubiquitous information access: Mobile hci 2003 international workshop, udine, italy. In *Lecture Notes in Computer Science*, volume 2954 / 2004, pages 143–157. Springer-Verlag Heidelberg, Sept 2003.
2. M. Swain and D. Ballard. Color indexing. *Intl. J. Computer Vision*, 7(1), 1991.