

SCENE RECOGNITION WITH CAMERA PHONES FOR TOURIST INFORMATION ACCESS

Joo-Hwee Lim, Yiqun Li, Yilun You, and Jean-Pierre Chevallet

French-Singapore IPAL Joint Lab (UMI CNRS 2955, I2R, NUS, UJF)
Institute for Infocomm Research, 21 Heng Mui Keng Terrace, Singapore 119613

ABSTRACT

Camera phones present new opportunities and challenges for mobile information association and retrieval. The visual input in the real environment is a new and rich interaction modality between a mobile user and vast information base connected to a user's device via rapidly advancing communication infrastructure. We have developed a system for tourist information access to provide scene description based on an image taken of the scene. In this paper, we describe the working system, the STOIC 101 database, and a new pattern discovery algorithm to learn image patches that are recurrent within a scene class and discriminative across others. We report preliminary scene recognition results on 90 scenes, trained on 5 images per scene, with an accuracy of 92% and 88% on a test set of 110 images, with and without location priming.

1. MOBILE VISUAL COMMUNICATION

Camera phones are becoming ubiquitous imaging devices: almost 9 out of 10 (89%) consumers will have cameras on their phones by 2009 (as forecasted by InfoTrends/Cap Ventures at www.capv.com). In 2007, camera phones will outsell all standalone cameras (i.e. film, single-use, and digital cameras combined). With this new non-voice non-text input modality augmented on a pervasive communication and computing device such as mobile phone, we witness emerging social practices of personal visual information authoring and sharing [1] and exploratory technical innovations to engineer intuitive, efficient, and enjoyable interaction interfaces [2].

In this paper, we present our study on using image input modality for information access in tourism applications. We describe a working system that provides multi-modal description (text, audio, and visual) of a tourist attraction based on its image captured and sent by a camera phone (Fig. 1). A recent field study [3] concludes that a significant number of tourists (37%) embraced the use of image-based object identification even when image recognition is a complex, lengthy and error-prone process. We aim to fulfill the strong desire of mobile tour guide users to obtain information on objects they come across during their visit, akin to pointing to a building or statue and asking a human tour guide "What's that?".

The AGAMEMNON project [4] also focuses on the use of mobile devices with embedded cameras to enhance visits



Fig. 1. Image-based mobile tour guide

of both archeological sites and museums. With the working rules that the input images are taken without or with minimum clutter, occlusion, and imaging variances in scale, translation, and illumination, a 95% recognition rate on 113 test images with 115 training images of only 4 target objects from 2 sites using mainly edge-based features has been reported.

The IDEixis system [5] is oriented towards using mobile image content with keywords extracted from matching web-pages to display relevant websites for user to select and browse. The image database was constructed from 12,000 web-crawled images where the qualities are difficult to control and the 50 test query images were centered around only 3 selected locations. The evaluation was based on image retrieval paradigm using the percentage of attempts their test subjects found at least one similar image among the first 16 retrieved images.

Our work differs from these systems that we focus on image recognition (instead of top similar matches) for significantly larger number of scenes (i.e. in the range of hundred) without making assumption on the input image. A key challenge in such image-based tour guide is the recognition of objects under varying image capturing conditions which is an open problem in computer vision research. Although 3D models provide a very powerful framework for invariant object representation and matching, it is very costly to build 3D models for large number of scenes, as compared to modeling of scenes through statistical learning from image examples that cover different appearances of the scenes.

In our current Snap2Tell prototype for tourist scene in-

formation access, we have developed a working system with Nokia N80 client, a unique database STOIC 101 (Singapore Tourist Object Identification Collection of 101 scenes), and a discriminative pattern discovery algorithm for learning and recognition of scene-specific local features. After describing the details of Snap2Tell in the next section, we describe superior experimental results on scene recognition when compared to state-of-the-art image matching methods that use global and local features.

2. THE SNAP2TELL SYSTEM

2.1. System Architecture

The Snap2Tell prototype is implemented as a 3-tier architecture: Client-Server-Database. The client is developed in J2ME on Nokia N80 and has functionalities to capture images and interact with the server. The client-server protocol is developed using XML for communication over WiFi and GPRS. Through the protocol, the client sends image queries and receives recognition results to and from the Java-based server respectively as depicted by the sequence of phone screen shots in Fig. 2. The server uses the recognition engine (c.f. 2.3) developed in C++ to identify the scene captured, retrieves and sends the scene descriptions, in both text and audio stored in Microsoft Access database, to the client.



Fig. 2. Screen shots of scene query and recognition

2.2. STOIC 101 Database

The STOIC 101 database consists of 101 Singapore tourist locations with a total of 5278 images. The images were taken at a proximate of 3 distances and 4 angles in natural light with a mix of occlusions and cluttered background to ensure a minimum of 16 images per scene. As illustrated in Fig. 3 (left), GPS coordinates had also been recorded at the circumference and as many user points of view as possible. Due to the unconstrained terrains, the other deciding factor would be taking photos at the angle most tourists would adopt (Fig. 3, right).

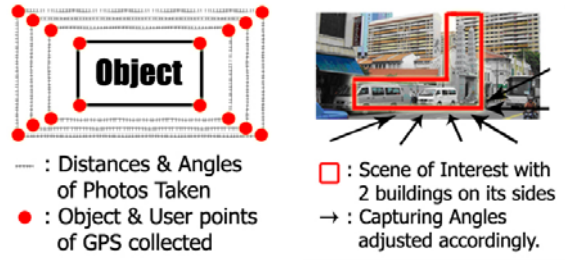


Fig. 3. Guideline for image and GPS collection

For every scene, its descriptions had been collected from various online sources. They were narrated into AMR audio format using Loquendo online Text-To-Speech engine (actor.loquendo.com). The wide spectrum of imaging conditions is to simulate unconstrained images taken by a casual tourist in real situations and makes the STOIC 101 database a challenging test collection for scene recognition. Fig. 4 depicts some sample images (two per scene).

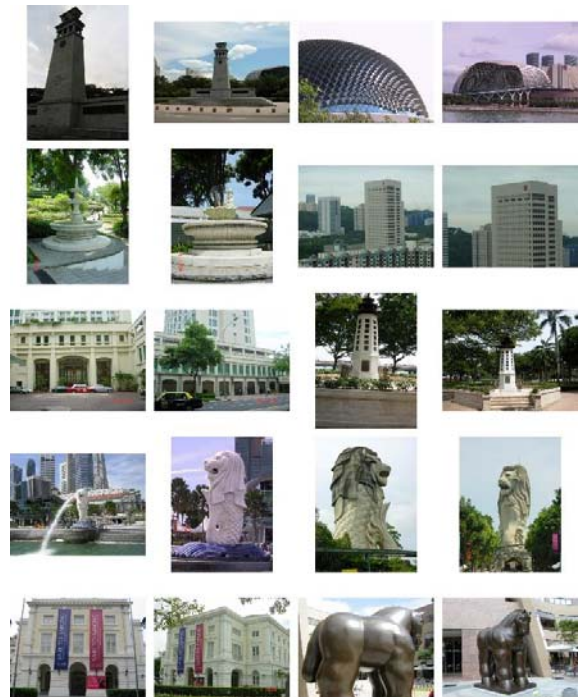


Fig. 4. Sample STOIC 101 database images

2.3. Scene Recognition using Discriminative Patches

Using invariant local descriptors of image patches extracted around interest points detected in an image for image matching and recognition is a very attractive approach. It represents a visual entity (object or scene) as its parts and allows flexible modeling of the geometrical relation among the parts. It can

focus on those parts in an image that are most important to recognize the visual entity for handling cluttered scenes with occlusions.

In terms of image representation, the “bag-of-visual-words” [6] scheme exploits the analogy between local descriptors in images and words in text documents. Training image patches from all classes are quantized (typically by k-means algorithm) into clusters to form a visual codebook. An image is then represented as histograms of cluster frequencies based on the image patches sampled in the image. However there are major problems with the unsupervised approach. Existing clustering methods favor high frequency patches which may not be more informative than patches with intermediate frequencies [7]. Furthermore, clusters of image patches suffer from polysemy and synonymy issues [6] i.e. as not all clusters have clear semantic interpretation!

In this paper, we propose a new pattern discovery approach to find local image patches that are recurrent within a scene class and discriminative across others. This selection strategy generates positive training patches for discriminative learning. We assume that when sufficient variety of scene classes is involved, the negative training samples are the union of the positive training examples for other classes. We use Support Vector Machines (SVMs) as the discriminative classifiers. We adopt multi-scale uniform sampling to extract patches from images instead of interest point scheme as the latter does not have advantage over random and uniform samplings when the sampling is dense enough [8].

2.3.1. Discriminative Patch Discovery

To discover discriminative patches, we compute the likelihood ratio for each image patch z sampled from the images,

$$L(z) = \frac{P(z|C)}{P(z|\bar{C})} \quad (1)$$

where C and \bar{C} are the positive and negative classes respectively. To estimate the likelihoods $P(z|C)$ and $P(z|\bar{C})$ from the patches in the training images of C and \bar{C} respectively, we can adopt the non-parametric density estimator such as Parzen-window [9].

As a rule of thumb, objects of interest in each class usually appear at the center of an image. For our experiments, we have designed a spatial weighting scheme to reward image patches near the center of the image as,

$$\omega(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}[(x_z - x_c)^2 + (y_z - y_c)^2]} \quad (2)$$

where x_z, y_z and x_c, y_c are the X-Y coordinates of patch z and image center respectively.

From our observation, the spatial weighting scheme has helped to select more relevant image patches. Thus we rank image patches by $\omega(z) \cdot L(z)$ and select the top image patches in a class as positive samples for that class and the union of

positive samples in all other classes as negative samples for that class. These automatically generated samples are then used to train local class-specific detectors using SVMs, denoted as $S_i(z)$ for each class C_i .

2.3.2. Discriminative Detection with Voting

Given a local patch sampled from an image, its visual features are computed and denoted as z . Then elements in the classification vector T for z can be normalized within $[0, 1]$ using the softmax function as

$$T_i(z) = \frac{\exp^{S_i(z)}}{\sum_j \exp^{S_j(z)}}. \quad (3)$$

In order to classify an image x (or recognize the object class i present in an image x), we aggregate the votes $V_i(x)$ of image patches z sampled from image x belonging to each class i as

$$V_i(x) = \sum_{z \in x} T_i(z), \quad (4)$$

and output the class which has the largest $V_i(x)$.

3. EXPERIMENTAL EVALUATION

We evaluate our new scene recognition approach on the STOIC 101 database. As a start, we report experiments based on 90 scene classes, each with 5 training images, and an independent test set of 110 images. For our experiments, we resized all images to 320×240 with both portrait and landscape layouts. Multi-scale patches ($60 \times 40, 80 \times 60, 100 \times 80, 160 \times 120$ with displacements of 40 and 30 pixels in horizontal and vertical directions respectively) are sampled for selecting top one-third of discriminative patches and later SVM learning with RBF kernels. Fig. 5 illustrates top discriminative patches identified on some sample images.

As we aim to support queries from large number of mobile users and to distribute computation such as feature extraction on the phones, we prefer features that allow efficient extraction whenever possible. For STOIC, we believe that color and edge features are most relevant to the image contents. Hence we did a lot of experiments with patch features such as linear color histograms in RGB, HSV or HS color channels (32 bins per channel), linear edge histograms (32 bins each for quantized magnitudes and angles), and combined color and edge histograms. The feature vectors are compared using simple city block distance metric. Color and edge features are combined linearly with equal weights in the SVM RBF kernel.

Table 1 lists selected results to show the effects of features and scales on recognition rates (C:color, E:edge). It is evident that color plays a dominant role though edge features and multi-scale sampling could improve the performance a little more with the best result of 88% using combined features of multi-scale patches.

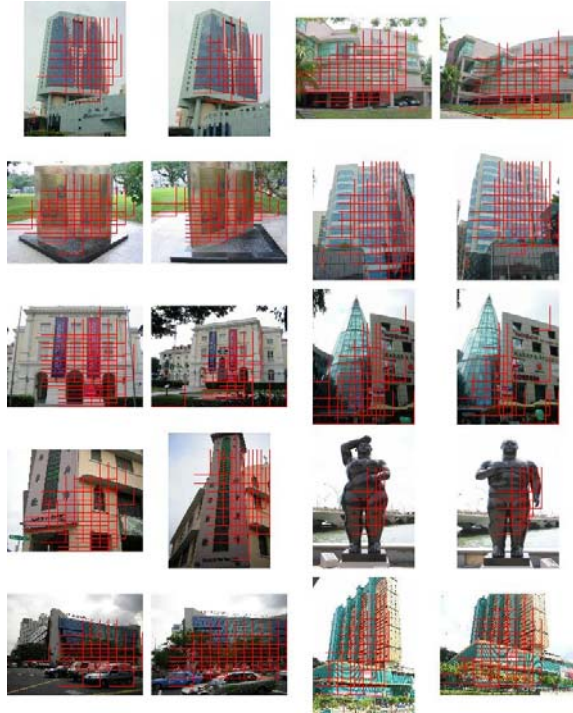


Fig. 5. Top discriminative patches for 10 scenes

With location priming to compare with only scenes in vicinity of a given query image (e.g. with the help of GPS coordinates), we have further increased the recognition rate to 92% which is significantly better than other methods as shown in Table 2: closest image matching using global histograms (C-H, CE-H), direct image matching based on keypoints using SIFT features (KP-G, KP-C), bags of visterms method (BoV) where 500 visterms are characterized by SIFT features and formed by k-means clustering, and visterm-based image signatures are used for SVM learning and classification, and our proposed method with and without pre-classification using location cues (DP, DP-L).

Table 1. Scene recognition results on STOIC subset

Features	Patch Sizes	# z	# Hit (%)
C	80×60	245	95(86%)
C	$80 \times 60, 100 \times 80, 160 \times 120$	550	92(84%)
C	$60 \times 40, 80 \times 60, 100 \times 80$	670	96(87%)
C + E	80×60	245	96(87%)
C + E	$60 \times 40, 80 \times 60, 100 \times 80$	670	97(88%)

4. CONCLUSIONS

In this paper, we proposed a tourist scene information access system using camera phone images. We have described

Table 2. Comparison with other methods

Notation	Methods	# Hit (%)
C-H	Color Histogram	84(76%)
CE-H	Color + Edge Histograms	85(77%)
KP-G	SIFT (grey) keypoint matching	78(71%)
KP-C	SIFT (color) keypoint matching	89(81%)
BoV	Bag of Visterms (SIFT)	68(62%)
DP	Discriminative Patches	97(88%)
DP-L	Discriminative Patches (localized)	101(92%)

the system architecture, the unique STOIC 101 dataset, and a scene learning and recognition algorithm based on pattern discovery that attains superior performance over several key global and local image matching methods. In the near future, we plan to perform a field trial with real tourists to evaluate the usability, efficiency, and recognition performance.

5. REFERENCES

- [1] D. Okabe and M. Ito, "Everyday contexts of camera phone use: steps towards technosocial ethnographic frameworks," in *Mobile Communication in Everyday Life*, J. Haflich and M. Hartmann, Eds. Berlin: Frank & Timme, 2006.
- [2] R. Ballags, J. Borchers, M. Rohs, and J.G. Sheridan, "The smart phone: a ubiquitous input device," *IEEE Pervasive Computing*, pp. 70–77, 2006.
- [3] N. Davies, K. Cheverst, A. Dix, and A. Hesse, "Understanding the role of image recognition in mobile tour guides," in *Proc. of MobileHCI*, 2005.
- [4] M. Ancona et al., "Mobile vision and cultural heritage: the agamemnon project," in *Proc. of 1st Intl. Workshop on Mobile Vision*, 2006.
- [5] K. Tollmar, T. Yeh, and T. Darrell, "Ideixis - image-based deixis for finding location-based information," in *Proc. of MobileHCI*, 2004.
- [6] P. Quelhas et al., "Modeling scenes with local descriptors and latent aspects," in *Proc. of IEEE ICCV 2005*, 2005.
- [7] F. Jurie and B. Triggs, "Creating efficient codebooks for visual recognition," in *Proc. of IEEE ICCV 2005*, 2005.
- [8] E. Nowak, F. Jurie, and B. Triggs, "Sampling strategies for bag-of-features image classification," in *Proc. of ECCV 2006*, 2006, pp. 490–503.
- [9] R.O. Duda, P.E. Hart, and D.G. Stock, *Pattern Classification*, Wiley, 2000.