

Identifying Different Settings in a Visual Diary

Michael Blighe^{1,2}, Noel E. O'Connor
Centre for Digital Video Processing
Dublin City University
Ireland
{blighem, oconnorn}@eeng.dcu.ie

Herwig Rehatschek, Gert Kienast
IIS³
Joanneum Research
Graz, Austria
{herwig.rehatschek, gert.kienast}@joanneum.at

Abstract

We describe an approach to identifying specific settings in large collections of photographs corresponding to a visual diary. An algorithm developed for setting detection should be capable of clustering images captured at the same real world locations (e.g. in the dining room at home, in front of the computer in the office, in the park, etc.). This requires the selection and implementation of suitable methods to identify visually similar backgrounds in images using their visual features. The goal of the work reported here is to automatically detect settings in images taken over a single week. We achieve this using Scale Invariant Feature Transform (SIFT) features and X-means clustering. In addition, we also explore how the use of location based meta-data can aid this process.

1 Introduction

A lot of research is currently taking place on the capture and retrieval of life logs in order to automatically generate a record of our daily lives [7] [1]. Much of the work focuses on using context and content information in order to infer details about one's daily activities [10]. Using photos, for example, one can easily construct a visual diary of an individual's life. For a single day, this might consist of a sequence of images providing a visual summary of the most important aspects of a person's day. The images used need to be selected from thousands of images representing an individual's day, and perhaps from millions over a lifetime. The key challenge is to manage, organise and search large vol-

umes of photos and to present representative samples in a visually coherent manner which is representative of events in that person's life.

In this work, we focus on personal image collections captured via a passive capture device, such as Microsoft's SenseCam [4]. We have developed an algorithm to perform *setting detection*. A *setting* in this context refers to those images taken at the same location in the real world (e.g. in the dining room at home, in front of the computer in the office, in the park, etc.).

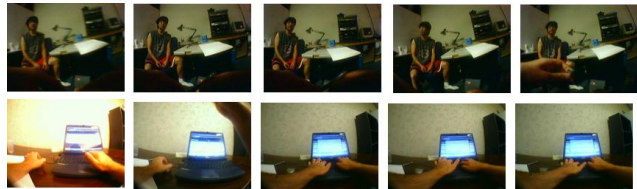


Figure 1. Images of two distinct settings

In order to achieve this, it is necessary to select and implement suitable methods to identify visually similar backgrounds in images using both visual and context based features. Our algorithm was developed using the SIFT features as they have proven their usefulness in a variety of object recognition tasks [8]. SIFT image features provide a set of features that are not affected by many of the complications experienced in other interest point detection methods, such as object scaling and rotation. Therefore, they provide an extremely useful method to detect similar objects in different SenseCam images, even if the background has been displaced or distorted. We captured location based data by logging Global System for Mobile Communication (GSM) signals. GSM potentially provides a ubiquitous source of location-based information. It works almost everywhere (i.e. indoors and outdoors), and requires no additional hardware to be carried by the user (besides their mobile handset).

The rest of this paper is organized as follows. In Sec-

¹The first author performed some of this work while at Joanneum Research

²The research leading to this paper was supported by the European Commission under contract FP6-045032 SEMEDIA, contract FP6-027026 (K-Space), Microsoft Research and Science Foundation Ireland under grant number 03/IN.3/I361.

³Institute of Information Systems & Information Management

tion 2, we review related research in this area. In section 3, we outline an approach to setting detection. Section 4 describes the experiments we performed and results obtained. In Section 5, we draw conclusions and outline future work.

2 Related Research

Many researchers have started work on developing passive capture devices - cameras which automatically take pictures without any user intervention. Gemmell et al [4] describe their work on the SenseCam, the device used in our work. They describe how passive capture lets people record their experiences without having to operate recording equipment, and without having to give recording a conscious thought. The advantages of this method of capturing photos are increased coverage of, and improved participation in, the event itself. However, the passive capture of photos presents new problems, particularly, how to manage and organise the massively increased volume of images captured.

Regarding scene detection, most works use colour and texture information to perform classification/retrieval. Vailaya et al. [15] used histograms of different low-level cues to perform scene *classification*. Different sets of cues were used depending on the two-class problem at hand: global edge features were used for city vs landscape classification, while local colour features were used in the indoor vs outdoor case. However, this approach is not really suitable for the detection of multiple settings and the use of colour means that their system is not very robust to viewing angle or lighting changes.

In order to provide ubiquitous coverage of a users location, we have two main choices - GSM or GPS. By 2010, mobile networks will cover 90% of the World's population [6]. In addition, mobile devices have long battery lives, constant connectivity and are nearly always at hand. This last point is extremely important, as it means the user doesn't have to carry around any additional sensors in order for their location to be tracked (unlike GPS). Recent studies have also shown that GPS coverage is only available for 4.5% of the time a user carries a device over a typical day [16]. In order to provide support for roaming, GSM mobile phones typically monitor six or seven neighbouring cells. This list of neighbouring cells will typically vary minimally when the mobile phone is static. However, the rate of change whilst moving will be more apparent, particularly in metropolitan environments with a large number of cells. Hence a change to neighbouring cells and signal strength levels typically indicates a change to the position of the mobile phone. Based on a pattern of observed cell towers and signal strengths, we can assume that users are in certain locations - although we need not necessarily require the ability to pin point the exact location.

3 Setting Detection

In the first step of our approach, the user reorganizes a single days SenseCam images to represent real settings. This is performed using a simple annotation tool which allows the user to update the setting information for each image. For any setting there are many features (i.e. interest points in the setting) that can be extracted to provide a *feature* description of the setting. This description can then be used when attempting to locate the setting in an image collection containing many other settings. In the case of SIFT descriptors, the extracted feature vector is a histogram. Each histogram's value corresponds to a weighted sum of the orientation of the images second moment matrix in a specific direction. Once the training data has been organised into distinct settings, we extract SIFT keypoints for each individual setting. SIFT keypoints are also extracted from each individual image in the test database.

3.1 Clustering

The X-means algorithm (an unsupervised variant of K-means) [12] is used to perform the clustering of the keypoints extracted from the settings selected from the training database. X-means is an extension of the K-means algorithm, where not only the position of the centres, but also the optimal number of clusters is estimated.

The algorithm starts with a small number of clusters (e.g. 3 centroids in Figure 2(b) below). Each cluster center is split in two child-points, to which the data-points belonging to the parent-point are distributed using a local K-means (with $k = 2$). The decision that has to be made is whether to replace the parent-point with the two child-points or not (i.e. the decision estimates whether the child points model the structure of the data better than the original parent model). The criterion we use to determine which model provides a better representation is the Bayesian Information Criterion although other kinds of probabilistic reasoning criteria might be used.

After clustering the keypoints for each test image using X-means, we then generate an image signature where m is the number of clusters, p_i is the center of the i^{th} cluster, and u_i is the relative size of the cluster (the number of descriptors in the cluster divided by the total number of descriptors extracted from the image [17]): $\{(p_1, u_1), \dots, (p_m, u_m)\}$

The Earth Mover's Distance (EMD) [13] is used to calculate the distance between signatures. It is defined as the minimum amount of work needed to change one signature into the other. The notion of work is based on a user-defined ground distance, which is the distance between two features. We use Euclidean distance as the ground distance. The EMD between two image signatures, $S1 : \{(p_1, u_1), \dots, (p_m, u_m)\}$ and $S2 : \{(q_1, w_1), \dots, (q_n, w_n)\}$, is de-

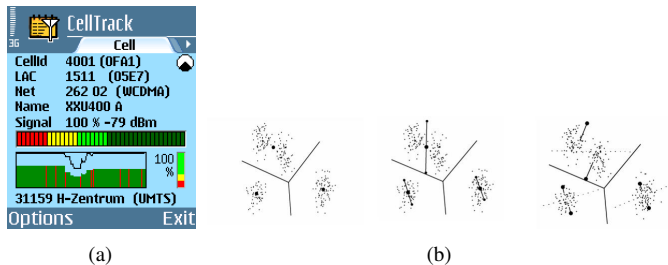


Figure 2. (a) CellTrack Interface; (b) Operation of X-means algorithm

fined as

$$D(S_1, S_2) = \frac{\sum_{i=1}^m \sum_{j=1}^n f_{i,j} d(p_i, q_j)}{\sum_{i=1}^m \sum_{j=1}^n f_{i,j}}$$

where $f_{i,j}$ denote a set of flows that minimize the overall cost and $d(p_i, q_j)$ is the ground distance between cluster centres p_i and q_j .

3.2 GSM Data Capture

We use CellTrack v1.18 [3] to capture GSM cell information (see figure 2(a)). This is a simple tool which runs on a Nokia Series 60 phone. The application logs numerous pieces of data which we can use to determine a users current location. Of particular interest are the *id* of the current cell, the *location area code* of the cell, the current network the phone is connected to, the signal strength and a user provided description of the current location. The application records these details every time there is a change of location, signal strength, etc. and the information is saved as a text file which can subsequently be exported from the mobile phone. We use the description provided by the user to label a particular setting.

The location based data was used to restrict the final settings the system could propose to a particular location. This was implemented by initially partitioning the database by location. The timestamps were used to match images to a particular location. So, for example, if there was only one setting at a particular location, we would not need to perform the analysis using the SIFT features as the location data alone would be sufficient to detect the setting in question. This situation did not arise in the data we use in this experiment. For the data selected in these experiments, we have two distinct locations. There are two settings at one location and four settings in the other.

4 Experimental Setup & Results

Two different experiments were performed involving a total of 14,965 images taken by the SenseCam over a pe-

riod of one week. The first days images of that particular week, 2,465 images in total, were used as training images. Using the annotation tool, these images were classified into different settings by the user. A total of six settings were found in that particular day. Each experiment was run using the procedure outlined above. However, in the second experiment, GSM data was included. The test images, 12,500 in total, were manually classified into different settings in order to provide a ground-truth for the experiment. A total of nine different settings were manually detected in the test images. Of these nine different settings, we only attempt to automatically detect the six settings we selected from the training data. This gave us a total of 8,299 relevant images from the test collection.

The results from the first experiment, along with the number of training and testing images in each setting, can be seen in Table 1. This shows the precision and recall figures for each setting for each experiment. All six settings were detected by the system. However, the recall and precision figures for certain settings vary considerably and some are quite low. In experiment 2, the contextual data, in the form of GSM location information, was introduced. This yielded a further improvement in the results across most settings. The location data allowed us to restrict the system to only propose settings in the same location as the test image. So, for example, a test image which contained a GSM Cell Id that indicated the user was in work, could only come from one of two settings - *Working on PC* or *At my desk*.

| Setting | Recall(1) | Precision(1) | Recall(2) | Precision(2) | Training images | Testing images |
|----------------|-----------|--------------|-----------|--------------|-----------------|----------------|
| Working on PC | 80.10% | 62.52% | 92.64% | 61.60% | 831 | 4334 |
| At home | 48.57% | 16.02% | 51.79% | 34.12% | 35 | 280 |
| Cooking dinner | 29.55% | 12.68% | 39.21% | 27.71% | 25 | 176 |
| At my desk | 9.12% | 52.37% | 13.87% | 55.90% | 502 | 2906 |
| Eating dinner | 73.57% | 13.39% | 82.14% | 52.27% | 18 | 140 |
| Reading in bed | 27.66% | 60.38% | 29.16% | 81.33% | 91 | 463 |

Table 1. Results from Experiment 1 & 2

5 Conclusions & Future Work

The main novelty of our work is the provision of a facility to browse a Visual Diary. An interesting result of performing setting detection is that once common settings have been determined (e.g. at work), more infrequent ones could be highlighted as they may be of more interest to the user. In addition, images which do not belong to any particular setting could also be highlighted to the user in a similar way. To the best of our knowledge, the method we have used has not been applied in this area before, although many variations on it certainly exist [17]. SIFT has been used successfully in many object detection systems, where the focus of the application is to detect specific objects within image scenes. In these applications, the training image often consists of only the particular object in question. With setting detection we are dealing with training images which can

contain multiple objects and cluttered backgrounds. The difficulty with this approach is that when we want to model a particular setting, there is no guarantee that the images we provide as training data adequately describe the underlying model. Furthermore, the fact that the user manually determines the training data, as well as ground truthing the test database, implies that a large element of subjectivity is introduced into the system. The limitations of this approach are clear in this work.

The X-means algorithm has been found to give good results when clustering SIFT keypoints [9] and, in general terms, it has been found to provide superior results to those obtained using a range of k values and the K-means algorithm [12]. However, the algorithm has several free parameters and the resulting number of clusters can vary greatly depending on the parameters used. In the current implementation, we believe that the number of clusters produced by X-means did not provide enough discriminative power to sufficiently model the settings in question. We intend to examine this and other clustering methods in the future. For example, in the first experiment, the X-means clustering returned 17 clusters for each setting. However, the first setting contained 324,937 keypoints and the fourth contained 10,022 keypoints. It seems clear that the total number of cluster centres used to create each setting signature was insufficient to generate good matching results. This assumption is further backed up when one examines the 1,000 cluster centres used to detect 7 different classes in [2]. In addition the volume of keypoints being used here, over 6,000,000 in these experiments, means that it may be difficult to find *distinctive* keypoints for matching an image to a particular setting. However, the great advantage of X-means is that it is unsupervised and we did not need to run the algorithm multiple times.

Much future work remains. The experiments highlight the value of using context based data when available to gain significant improvements. This type of data is readily available due to the abundance of sensors on the SenseCam itself and by using other information available from a mobile phone. We believe the choice of SIFT is justified due to the wealth of information in the object detection and recognition literature [11] [5]. However, a comprehensive evaluation of a number of interest point detection algorithms will be undertaken. A similar study has been previously performed [14], however, the distortion and other image changes were manually added to the images. We plan to perform an evaluation using the SenseCam images themselves where all these changes occur naturally.

References

- [1] M. Blighe, H. L. Borgne, and N. O'Connor. Exploiting context information to aid landmark detection in sensecam im-

- ages. In *2nd International Workshop on Exploiting Context Histories in Smart Environments - Infrastructures and Design (ECHISE)*, September 2006.
- [2] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray. Visual categorization with bags of keypoints. In *European Conference on Computer Vision*, May 2003.
- [3] A. Fischer. Celltrack. <http://www.afischer-online.de/sos/celltrack/>.
- [4] J. Gemmell, L. Williams, K. Wood, R. Lueder, and G. Bell. Passive capture and ensuing issues for a personal lifetime store. October 2004.
- [5] K. Grauman and T. Darrell. Efficient image matching with distributions of local invariant features. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, page 627634, 2005.
- [6] GsmWorld. Mobile coverage by 2010. <http://www.gsmworld.com/news/headlines.shtml>.
- [7] M. Lamming and M. Flynn. Forget-me-not: Intimate computing in support of human memory. In *FRIEND21, Int. Symp. Next Generation Human Interface*, pages 125–128, February 1994.
- [8] D. Lowe. Distinctive image features from scale-invariant keypoints. In *International Journal of Computer Vision*, volume 60(2), pages 91–110, 2004.
- [9] A. Noulas and B. J. A. Krse. Unsupervised visual object class recognition. In *Advanced School of Computing and Imaging Conference*, Lommel, Belgium, 2006.
- [10] N. O'Hare, C. Gurrin, G. Jones, and A. Smeaton. Combination of content analysis and context features for digital photograph retrieval. *2nd IEE European Workshop on the Integration of Knowledge, Semantic and Digital Media Technologies*, November 2005.
- [11] A. Opelt, M. Fussenegger, A. Pinz, and P. Auer. Weak hypotheses and boosting for generic object detection and recognition. In *European Conference on Computer Vision*, volume 2, page 7184, 2004.
- [12] D. Pelleg and A. Moore. X-means - extending k-means with efficient estimation of the number of clusters. In *17th International Conference on Machine Learning*, pages 727–734, 2000.
- [13] Y. Rubner, C. Tomasi, and L. Guibas. The earth mover's distance as a metric for image retrieval. In *International Journal of Computer Vision*, volume 40(2), pages 99–121, 2000.
- [14] C. Schmid, R. Mohr, and C. Bauckhage. Evaluation of interest point detectors. *International Journal of Computer Vision*, 37(2):151–172, 2000.
- [15] A. Vailaya, M. Figueiredo, A. Jain, and H. Zhang. Image classification for content-based indexing. In *IEEE Trans. on Image Processing*, volume 10 of *I*, pages 117–130, 2001.
- [16] A. Varshavsky, M. Chen, E. de Lara, J. Froehlich, D. Haehnel, J. Hightower, A. LaMarca, F. Potter, T. Sohn, K. Tang, and I. Smith. Are gsm phones the solution for localization? In *7th IEEE Workshop on Mobile Computing Systems and Applications (HotMobile)*, 2006.
- [17] J. Zhang, M. Marszalek, S. Lazebnik, and C. Schmid. Local features and kernels for classification of texture and object categories: An in-depth study. Technical Report RR-5737, INRIA Rhone-Alpes, November 2005.