# IDENTIFYING DIFFERENT SETTINGS IN A VISUAL DIARY

*Michael Blighe*[*][†], *Noel E. O'Connor*

Centre for Digital Video Processing,
Dublin City University,
Ireland

*Gert Kienast, Herwig Rehatschek*

Institute of Information Systems & Information Management,
Joanneum Research,
Graz, Austria

## ABSTRACT

In this paper, 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 preliminary work reported here is to semi-automatically detect settings in SenseCam images taken over a single day. We achieve this via a user driven clustering process that uses low-level MPEG-7 and Scale Invariant Feature Transform (SIFT) features.

## 1. INTRODUCTION

A vast amount of largely un-indexed, heterogeneous data, currently exists in professional and personal media repositories and Intranets all over the world. A key problem faced when managing this type of data is how to organise, search and retrieve it. It is clear that new methods, environments and widely usable tools for media labelling, searching and retrieval from very large collections of heterogeneous data are needed. These methods include building on and extending research in media technologies, web semantics, artificial intelligence, content based image retrieval and interface design.

In this work, we focus on personal collections, and specifically on large collections of still images that constitute an individual visual diary, such as those captured using Microsoft's SenseCam (see figure 1). We have developed an algorithm to perform semi-automatic 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.).

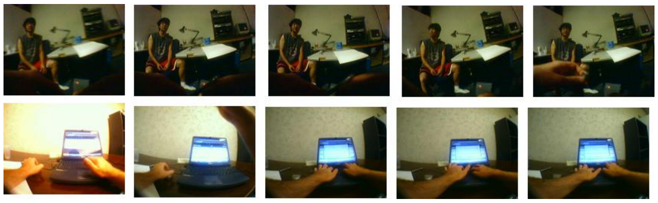

**Fig. 1**. Sample SenseCam images showing two distinct settings

In order to achieve this, it is necessary to select and implement suitable methods to identify visually similar backgrounds in images using (in this case) only visual features. Our algorithm was developed using the SIFT features as they have proven their usefulness in a variety of object recognition tasks [1]. 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.

## 2. SETTING DETECTION

In our semi-automatic approach, we first cluster a set of training images using a simple K-means algorithm and standard MPEG-7 features. This is performed in order to facilitate a subsequent re-organization of these clusters to represent real settings by the user. Given this user generated training data, we extract signatures for each setting and then classify test images accordingly. The objective is to provide the user with a low-overhead mechanism for organising his/her visual diary in terms of specific settings of interest. Low-level MPEG-7 features were extracted from the images using the aceToolbox [2]. The descriptors used were Colour Layout, Scalable Colour, Colour Structure and Edge Histogram. Different clustering algorithms were used to cluster both the low-level features and the SIFT keypoints extracted from the images. In addition, a basic annotation tool was developed to allow users to update the initial cluster information generated by the system. Each of these steps is described in more detail below.

[*] The first author performed some of this work while at Joanneum Research

[†] The research leading to this paper was supported by the European Commission under contract FP6-027026 (K-Space), the aceMedia Project under contract FP6-001765, Microsoft Research and Science Foundation Ireland under grant number 03/IN.3/I361.

### 2.1. CLUSTERING

The K-means algorithm was used to perform an initial clustering of the training images using the low-level MPEG-7 features extracted from each image. Clustering of the SIFT keypoints extracted from each test image on the other hand was performed using the x-means algorithm (an unsupervised variant of k-means) [3]. X-means is an extension of the k-means algorithm, where not only the position of the centers, but also the optimal number of clusters is estimated.

### 2.2. Image Signature

After clustering the keypoints for each test image using x-means, we save the cluster centres. We then generate an image signature where *m* is the number of clusters, $p_i$ is the center of the *ith* 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 [4]): $\{(\mathrm{p}_1,\mathrm{u}_1),...,(\mathrm{p}_m,\mathrm{u}_m)\}$.

### 2.3. Earth Mover's Distance

The Earth Mover's Distance (EMD) [5] 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*:$\{(\mathrm{p}_1,\mathrm{u}_1),...,(\mathrm{p}_m,\mathrm{u}_m)\}$ and *S2*:$\{(\mathrm{q}_1,\mathrm{w}_1),...,(\mathrm{q}_n,\mathrm{w}_n)\}$, is defined below. We want to find a set of flows, $\mathrm{f}_{i,j}$, that minimize the overall cost where $\mathrm{f}_{i,j}$ is a flow value that can be determined by solving a linear programming problem and $\mathrm{d}(\mathrm{p}_i, \mathrm{q}_j)$ is the ground distance between cluster centres $\mathrm{p}_i$ and $\mathrm{q}_j$.

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

## 3. EXPERIMENTS & RESULTS

A total of 1,318 images (659 training & 659 test) taken by the SenseCam on a single day were used in this simple experiment. The MPEG-7 features were concatenated together to create one representative vector for each training image. These were then clustered using the K-means algorithm with k=5. Using the annotation tool, the user then inspected each of these clusters and manually updated the cluster information in order to create clusters which represented distinct *settings*. For each of the updated clusters, the SIFT keypoints were extracted for their image members and clustered using the X-means algorithm. The cluster centres were saved and the image signature created. This gave us seven signatures which were representative of seven individual settings from the image collection.

The next step was to take the keypoints for each test image individually and cluster the extracted SIFT keypoints using X-means. Again, the cluster centres for each image were saved and a signature generated for each image. Earth Mover's Distance was used to calculate the distance between each image signature and each of the seven cluster signatures. Where the distance was minimal, the image was deemed to belong to that particular cluster and hence that setting. In order to evaluate the results, the final clusters generated by this process were compared to settings selected by the user to determine if they matched. Out of a total of 659 test images, a total of 325 (49.32%) matched the settings selected by the user.

## 4. CONCLUSIONS

This paper demonstrates the potential of using SIFT keypoints to perform setting detection in SenseCam images. However, the results generated in this simple experiment demonstrate the need for additional information in order to improve results. In this regard, we plan to use low-level MPEG-7 features, location based data, and sensor information taken from the SenseCam. In addition, we will substantially increase the amount of data used in the experiments and also look at different clustering and dimensionality reduction techniques in order to improve performance. A more comprehensive evaluation framework will also be developed based around an updated version of the annotation tool.

## 5. REFERENCES

[1] D. Lowe, "Distinctive image features from scale-invariant keypoints," in *International Journal of Computer Vision*, 2004, vol. 60(2), pp. 91–110.

[2] N. OConnor, E. Cooke, H. Le Borgne, M. Blighe, and T. Adamek, "The acetoolbox: Low-level audiovisual feature extraction for retrieval and classification," *2nd IEE European Workshop on the Integration of Knowledge, Semantic and Digital Media Technologies*, November 2005.

[3] 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*, 2000, pp. 727–734.

[4] J. Zhang, M. Marszalek, S. Lazebnik, and C. Schmid, "Local features and kernels for classifcation of texture and object categories: An in-depth study," Tech. Rep. RR-5737, INRIA Rhone-Alpes, November 2005.

[5] Y. Rubner, C. Tomasi, and L. Guibas, "The earth mover's distance as a metric for image retrieval," in *International Journal of Computer Vision*, 2000, vol. 40(2), pp. 99–121.