Exploiting Context Information to aid Landmark Detection in SenseCam Images*

Michael Blighe[1], Hervé Le Borgne[1], Noel E. O'Connor[1,2], Alan F. Smeaton[1,2] and Gareth J.F. Jones[1]
[1]Centre for Digital Video Processing and [2]Adaptive Information Cluster,
Dublin City University,
Ireland
blighem@eeng.dcu.ie

ABSTRACT

In this paper, we describe an approach designed to exploit context information in order to aid the detection of landmark images from a large collection of photographs. The photographs were generated using Microsoft's SenseCam, a device designed to passively record a visual diary and cover a typical day of the user wearing the camera. The proliferation of digital photos along with the associated problems of managing and organising these collections provide the background motivation for this work. We believe more ubiquitious cameras, such as SenseCam, will become the norm in the future and the management of the volume of data generated by such devices is a key issue. The goal of the work reported here is to use context information to assist in the detection of landmark images or sequences of images from the thousands of photos taken daily by SenseCam. We will achieve this by analysing the images using low-level MPEG-7 features along with metadata provided by SenseCam, followed by simple clustering to identify the landmark images.

Keywords

SenseCam, context, low-level features, clustering, landmark images

1. INTRODUCTION

The management of personal collections of digital photos is becoming an increasingly difficult task. The proliferation of digital cameras and cameraphones means that taking pictures has never been easier. Gradually, we are getting closer to Vannevar Bush's 1945 Memex vision [3] of storing a lifetime's worth of documents and photographs. However, the usefulness of the collected photos is dubious, as the methods of managing digital photos have not kept pace with the technology for acquiring them. Naaman et al [11] describe how the photo collection management problem can be categorized into tools which enable easy annotation of photos, tools which allow fast visual scanning of the images and content-based tools. However, they also identify the problems associated with each of these types of systems, such as difficulties for consumers with annotation, inability of tools to allow fast visual scanning to scale to many thousands of images and the semantic gap in relation to content based tools.

In future, digital cameras will become more ubiquitious, and they will eventually be integrated into all facets of our daily lives. This will only serve to exacerbate the problems with current photo management systems. Already, researchers have started work on passive capture devices - cameras which automatically take pictures without any user intervention. Gemmell et al describe their work on the SenseCam, the device used in our work, in [6]. 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, and improved participation in the event itself. Healey et al [8] describe a system called StartleCam which is a wearable video camera, computer, and sensing system which also passively captures images depending on certain events detected by the sensors on the device.

However, the passive capture of photos presents new problems, particularly, how to manage and organise the massively increased volume of images captured. Traditional systems for content-based image retrieval are not adequate for this task. In [5] the authors describe the MyLifeBits system, which is a first step in tackling this problem, specifically in relation to the images captured by SenseCam. MyLifeBits also captures other forms of digital media and is a step towards fulfilling Bush's Memex vision.

The use of automatically collected metadata has been shown to be helpful in the organization of photo collections. The use of camera metadata, in combination with low-level features, is discussed in [2], where a Bayesian network is used to fuse content-based data and metadata, with some promising results in specific contexts (e.g. indoor/outdoor classification).

One approach which may be useful in tackling this problem is to exploit context and in particular context histories. Wolf

*

et al [14] describe how a dichotomy of useful information sources exist to detect objects within an image: appearance and context. They describe how appearance information includes patterns of brightness, edge responses, color histograms, texture cues, and other features commonly used for object detection, whereas contextual-features are somewhat more loosely defined. They define context as *information relevant to the detection task but not directly due to the physical appearance of the object* and also describe how contextual metadata can serve as a memory cue and can also imply the content of the image. Other authors who have empoyed context and content analysis to aid photograph and video retrieval include [13] & [1].

In our work with SenseCam, we plan to use a combination of low-level content analysis and metadata provided by the SenseCam itself, to generate landmark images for a single day's SenseCam photographs. The ultimate goal is to generate a context history to allow us to usefully infer further information from a number of days photos and to link landmarks from different day's photographs together.

The rest of this paper is organized as follows. In Section 2, we provide a high level description of the low-level features and metadata used in the experiments, along with a description of what constitutes a landmark image in the context of our work. Section 3 describes the algorithm used to perform clustering and to detect the landmark images. Section 4 describes the experiment we performed, whilst Section 5 outlines the results obtained. Future work and our vision for the future of this system are discussed in Section 6 and some conclusions are drawn in Section 7.

2. LOW-LEVEL FEATURES & CONTEXT DATA

A number of MPEG-7 low-level features, along with metadata taken directly from the SenseCam, were used in this experiment. A brief description of these follows.

2.1 MPEG-7 features

The aceToolbox was used to extract low-level features from the SenseCam images. A more detailed description of the aceToolbox can be found in [12]. A brief description of the descriptors used in this experiment is provided here and more detailed information can be found in [10].

- Scalable Colour generates a colour histogram in the hue saturation value (HSV) colour space that is encoded using a Haar transform thereby providing a scalable representation.
- Colour Layout is designed to capture the spatial distribution of colour in an image or region by clustering the image into 64 blocks and deriving the average colour of each block. These values are then transformed into a series of coefficients by performing an 8×8 DCT.
- The Edge Histogram captures the spatial distribution of edges, which are identified using the Canny algorithm, by dividing the image into 16 non-overlapping blocks and then calculating 5 edge directions in each block.
- Homogeneous Texture describes directionality, coarseness and regularity of patterns in images by partitioning the images frequency domain into 30 channels and computing the energy and energy deviation of each channel and outputting the mean and standard deviation of the frequency coefficients.

2.2 SenseCam

Detailed technical information about SenseCam can be found in [6]. We use version 2.3 of the SenseCam shown in Figure 1 along with an image of the device in Figure 2. The SenseCam takes pictures automatically by default every fifty seconds. It also has a number of sensors onboard the device which trigger capture more frequently. The sensors include a passive infra-red detector (similar to that used in home alarm systems) which can detect people or other warm objects directly in front of the individual wearing the camera, an accelerometer which captures data in the X, Y & Z directions, a digital light sensor and a temperature sensor. We propose to use the temperature, light and accelerometer values in this experiment. In a typical day, SenseCam will record anything between 2,000 and 3,000 photos.

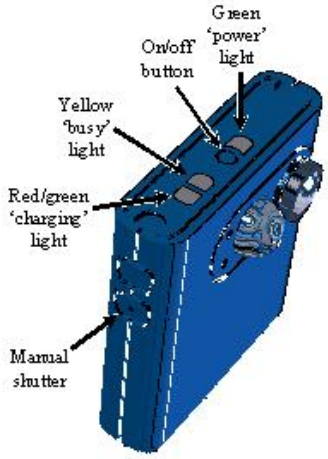

Figure 1: Schematic of Microsoft SenseCam

Figure 2: Microsoft SenseCam

2.3 Landmark Images

A landmark can be defined as a prominent, identifying, feature of a landscape. Normally, it is associated with a single object or place in the landscape. In our work, we define a landmark to be a single image, or a number of images, which are temporally aligned. In terms of the images produced by SenseCam on a daily basis, a landmark will be an image, or images, which represent a particular event in an individual's day. Our initial analysis is restricted to SenseCam images from one day. An event in this context could be simply defined as the elements that make up a persons day. For example, in the morning the images taken when one gets out of bed, prepares for work, and has breakfast could be considered an event. Another event would be when the individual leaves the home to travel to work. The goal then for this particular experiment is to try and detect these events and to extract landmark images to represent them. Ultimately, we would hope to generate a context history, using the low-level features and metadata described above, to assist us in generating landmarks over a period of days.

In Figure 3, an artificial graphical illustration of what a landmark image might be is given. The image depicts the temperature, light sensor and accelermoter readings from the SenseCam, as well as the combined low-level feature vector, over a period of one day. Key points where both sensors increase or decrease suddenly can be interpreted as being event boundaries (i.e. a significant change from one event to another within the days images). Other information might also be usefully inferred from these readings. In the scenario below, the accelerometer values increase sharply while all other readings are falling. An analysis of these trends within the data may reveal useful contextual information which can aid detection. A landmark image could then be selected from within each event. For example, an initial approach to this would be to simply select the middle image within each event. However, more advanced methods will be explored in future in order to choose landmark images which are more representative of the detected events.

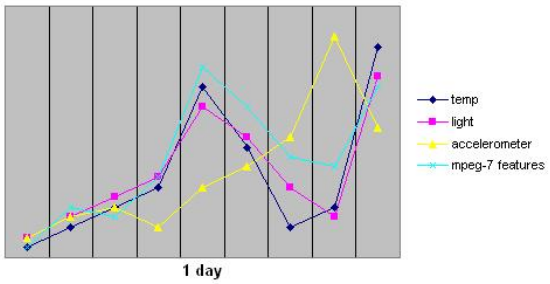

Figure 3: Landmark Detection over one day

3. CLUSTERING & DETECTION

As an initial step in this work, a simple clustering algorithm has been employed to form clusters and to detect a set of landmark images for the entire day's collection. No event boundary analysis has been employed in the experiments reported here.

The algorithm used in based on the agglomerative hierarchical clustering approach [4]. This approach starts by considering each individual photograph as a cluster, and the sequence is then formed by successively merging clusters. The merging is performed based on the nearest distance between photographs, where the distance calculated is the Euclidian distance based on a feature vector containing the normalised low-level features and metadata for each image. The data was normalised between values of 1 and 0. Time constraints are also imposed on the clustering process based on an algorithm proposed in [9]. This is implemented by considering the time each photo was taken and penalising photos taken further away from each other (in time) using a cost function, thus increasing the distance measure. The cost function is calculated based on the average squared distance of the data set.

Once the merging process has been completed, a dendrogram [4] is created which graphically illustrates how the individual images have been clustered. Due to the large number of images used in this experiment, over 2000 per day, it is not feasible to view the entire dendrogram as it contains too much information. Instead, we will view the dendrogram at different offsets from the top of the tree. Figure 4 shows a dendrogram containing the top 20 nodes.

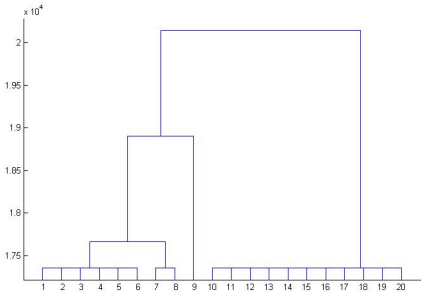

Figure 4: Dendrogram containing top 20 nodes

By analysing the tree at different levels, we can inspect the clusters at that level and select a single landmark image from within each cluster for that particular day. This type of analysis is interesting in the context of the SenseCam as we are more interested in getting a representative set of landmark images at different times during the day than in providing one single representative image for the whole day. By selecting different cutoff points at different levels of the tree, we can vary the amount of landmark images we wish to select as the final result.

4. EXPERIMENT

The set of images produced by SenseCam taken on the 9th June 2006 by the first author contained typical events from a normal day, and also some unusual events. The number of photos used in the experiment was 2,243 with the first photograph being taken at 08:34 in the morning and the last photograph taken at 21:41 in the evening. The aceToolbox was used to generate the MPEG-7 features outlined above and the results for each low-level feature were concatenated together to create one representative vector for each image. The accelerometer, temperature and light sensor readings were extracted from the CSV file produced by SenseCam. All data was normalised before the Euclidian distance was

calculated. The distance was calculated separately for the low-level features and for each of the metadata features giving 4 distance measures for each image. The results of these were then concatenated together. Time constraints were implemented by extracting the time each photo was taken from the same CSV file. As this analysis was being undertaken within the context of a single day, these times were converted into the number of seconds from midnight. The cost function was then calculated and the time constraints imposed on the distance matrix, giving a final matrix, 2433×2433, containing the distances between each image and all the others in the database. The clustering approach outlined above was then used to generate the landmark images for the day's collection.

Finally, in order to allow for an evaluation of the results, the images for this particular day were manually segmented into events (as defined above) and these events can be seen in the table below. Landmark images were not manually chosen for each event in this particular experiment, as the landmarks extracted are simply the middle image of each event. By examining the results obtained in this experiment, we can see whether the landmarks chosen represent the manually annotated events. This provides us with an initial evaluation of our work.

Manually Annotated Events
Morning time at home
Leaving home to cycle to the park
Meeting a friend in the park
Cycling and chatting in park
SenseCam covered by cycling jersey
At home in the afternoon
Leaving home and going shopping
At home in the evening
SenseCam taken off and pointing out to balcony
At home watching TV and having dinner

Table 1: Manually annotated events

5. RESULTS

The results presented show landmark images selected from dendrograms containing 5, 10, 15 and 20 nodes. These values were selected at random and any number of nodes could have been selected in order to view results at that particular level. The thumbnails in figures 5, 6, 7 and 8 show the landmark images selected at each of the offsets above.

An initial analysis of the results shows that all events have been recognised in the sequence of landmark images shown in figure 5. For example, the first 3 images on the top row represent the first event *Morning time at home*. The second, third and fourth images on the second row represent the second, third and fourth events and the last two images on the third row along with the first two images on the fourth row represent the event *Leaving Home and going shopping*. In figures 6, 7 and 8, we lose certain events as we reduce the number of images being presented to the user. This is an expected outcome of this method of displaying a flexible result set. The completely black image contained in figures 5, 6 and 7 represents the *SenseCam covered by cycling jersey* event. At this particular point in the day, the user had placed the SenseCam underneath a cycling jersey without

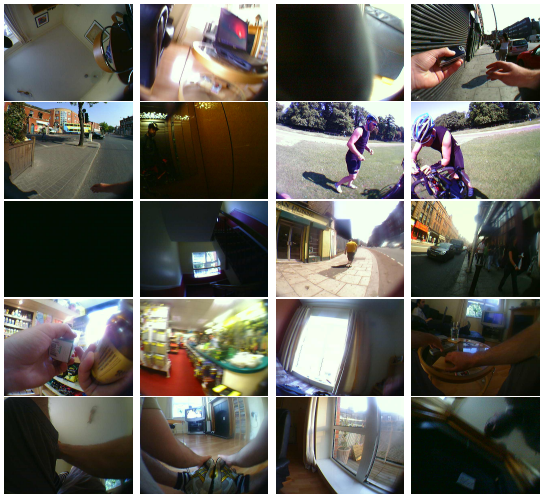

Figure 5: Landmark Images from top 20 nodes

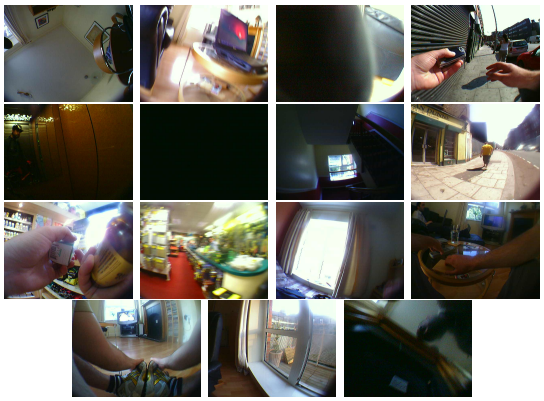

Figure 6: Landmark Images from top 15 nodes

turning the device off. These images are quite distinctive, so one would expect them to have been recognised as a seperate event. We also note the presence of some blurred images which, although representative of a particular event, are not the best images to display as a final result. However, the nature of the SenseCam application environment means that although these images are not the best photographs to display to another user, they should prove meaningful to the individual who wore the SenseCam while the photographs were being taken.

As a first step, we believe our results show that the simple approach used in this work demonstrates the potential benefits of using low-level features and context information to generate a representative collection of landmark images from the SenseCam data.

6. FUTURE WORK

Future work will focus on the refinement of the algorithm outlined in this work, in order to improve results. The method employed to detect landmark images within a single day's SenseCam photographs will be improved by exploring more intelligent clustering methods, including new methods of imposing time constraints, as well as exploring the use of

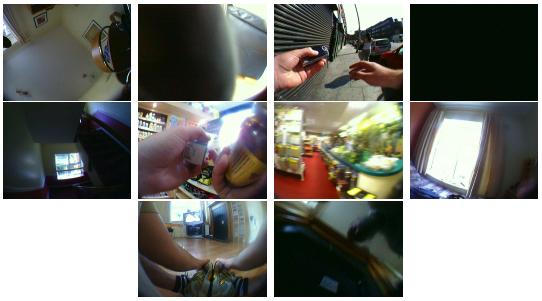

Figure 7: Landmark Images from top 10 nodes

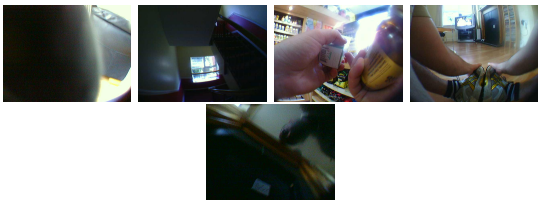

Figure 8: Landmark Images from top 5 nodes

other low-level MPEG-7 features available within the aceToolbox. We will also focus on more sophisticated methods of fusing the low-level features and SenseCam metadata in order to improve results.

We are also currently gathering SenseCam images and correlating with external sources of metadata. A single user wearing a SenseCam can also wear a heart-rate monitor and a BodyMedia device and carry a GPS device in order to provide a richer metadata set to explore. Again, the possibilites to incorporate this extra metadata into our application to improve results will be explored. We also plan to investigate partial image analysis, perhaps using segmentation or salience (using Harris point detectors [7]), or even face detection techniques in order to acquire improved results.

7. CONCLUSIONS

This paper has demonstrated an initial approach to using context in helping to detect landmark images from within the several thousands of images which make up a single day's SenseCam photographs. The approach is simple, but works well, and can efficiently highlight events and images of interest from an individual's day. We have started to use some simple context, namely the temporal aspect and SenseCam metadata, for our landmark detection but there is a large range of correlated metadata which we can also incorporate into the process. This paper also outlines our vision for future work in the area with SenseCam images and we have outlined our vision for an application exploiting context histories within the SenseCam application environment. This future work continues.

Acknowledgements

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.

8. REFERENCES

[1] K. Aizawa, D. Tancharoen, S. Kawasaki, and T. Yamasaki. Efficient retrieval of life log based on context and content. *CARPE*, October 2004.

[2] M. Boutell and J. Luo. Beyond pixels: Exploiting camera metadata for photo classification. *IEEE Conference on Computer Vision and Pattern Recognition*, June 2004.

[3] V. Bush. *As we may think*. The Atlantic Monthly, July 1945.

[4] R. Duda, P. Hart, and D. Stork. *Pattern Classification*. Wiley, 2001.

[5] J. Gemmell, A. Aris, and R. Lueder. Telling stories with mylifebits. In *IEEE International Conference on Multimedia Expo*, pages 1536–1539, July 2005.

[6] J. Gemmell, L. Williams, K. Wood, R. Lueder, and G. Bell. Passive capture and ensuing issues for a personal lifetime store. In *1st ACM Workshop on Continuous Archival and Retrieval of Personal Experiences*, pages 48–55. ACM, October 2004.

[7] C. Harris and M. Stephens. A combined corner and edge detector. In *4th Alvey Vision Conf.*, pages 147–151, 1988.

[8] J. Healey and R. Picard. Startlecam: A cybernetic wearable camera. In *Second International Symposium on Wearable Computing*, pages 42–49, Pittsburgh, PA, October 1998.

[9] W. Lin and A. Hauptmann. Structuring continuous video recordings of everyday life using time constrained clustering. *SPIE Symposium on Electronic Imaging*, January 2006. San Jose, CA.

[10] B. Manjunath, J.-R. Ohm, V. Vasudevan, and A. Yamada. Color and texture descriptors. *IEEE trans. circuits and systems for video technology*, 11(6):703–715, 2001.

[11] M. Naaman, S. Harada, Q. Wang, H. Garcia-Molina, and A. Paepcke. Context data in geo-referenced digital photo collections. In *12th ACM International Conference on Multimedia*, pages 196–203, October 2004.

[12] N. OConnor, E. Cooke, H. L. 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.

[13] 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.

[14] L. Wolf and S. Bileschi. A critical view of context. *International Journal of Computer Vision*, 69(2):251–261, April 2006.