Multimodal Segmentation of Lifelog Data

Aiden R. Doherty[1], Alan F. Smeaton[1], Keansub Lee[2] & Daniel P.W. Ellis[2]
Centre for Digital Video Proessing & Adaptive Information Cluster, Dublin City University, Ireland[1]
LabROSA, Columbia University, New York, USA[2]
{adoherty, asmeaton}@computing.dcu.ie, {dpwe, kslee}@ee.columbia.edu

Abstract

A personal lifelog of visual and audio information can be very helpful as a human memory augmentation tool. The SenseCam, a passive wearable camera, used in conjunction with an iRiver MP3 audio recorder, will capture over 20,000 images and 100 hours of audio per week. If used constantly, very soon this would build up to a substantial collection of personal data. To gain real value from this collection it is important to automatically segment the data into meaningful units or activities. This paper investigates the optimal combination of data sources to segment personal data into such activities. 5 data sources were logged and processed to segment a collection of personal data, namely: image processing on captured SenseCam images; audio processing on captured iRiver audio data; and processing of the temperature, white light level, and accelerometer sensors onboard the SenseCam device. The results indicate that a combination of the image, light and accelerometer sensor data segments our collection of personal data better than a combination of all 5 data sources. The accelerometer sensor is good for detecting when the user moves to a new location, while the image and light sensors are good for detecting changes in wearer activity within the same location, as well as detecting when the wearer socially interacts with others.

Introduction

The SenseCam, developed by Microsoft Research Cambridge, is a small wearable device which incorporates a digital camera and multiple sensors including: sensors to detect changes in light levels, an accelerometer to detect motion, a thermometer to detect ambient temperature, and a passive infra red sensor to detect the presence of a person. Sensor data is captured approximately every 2 seconds and based on these readings it is determined when an image should be captured. For example the light sensor will trigger the capture of an image when the wearer is moving between two different rooms as there will be a distinct change in the light level as the wearer moves towards the door, opens it and moves into a new room. An image is also captured when the passive infrared sensor detects the presence of a person arriving in front of the device indicating that the wearer has just met somebody. The accelerometer sensor is useful as a lack of motion indicates an optimal time to take a non-blurred image. If there is no image captured based on sensor activity over a predetermined time period (50 seconds), an image will be automatically captured. All the sensor data is correlated with the captured SenseCam images when downloaded to a computer.

The iRiver T10 is an MP3 player with 1GB of flash memory and is powered by a single AA battery. This device also has a built-in microphone and is capable of recording MP3 data sampled at 64 kbps, which means that an entire day's worth of data can be easily recorded. Figure 1 depicts an individual wearing the SenseCam in front of his chest, via a strap around his neck, and an audio recorder clipped on to the right belt strap of his trousers.

Hodges et. al. (2006) detail the potential benefits of a personal visual diary such as that generated by a SenseCam or audio recorder. In preliminary experiments they have found that the use of a SenseCam dramatically aided a subject, suffering from a neurodegenerative disease (limbic encephalitis), to recall events that happened during her day when reviewing that day's

activities using SenseCam images. From personal experience the authors of this paper have also found improved short-term memory recall of activities experienced during days while wearing the SenseCam device.

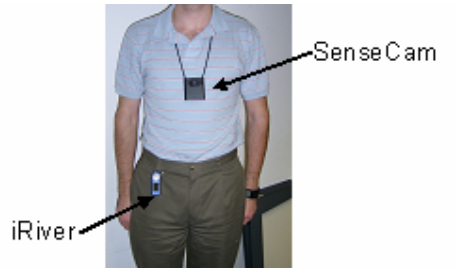

Figure 1 SenseCam and iRiver audio recoder

A SenseCam captures 3,000 images on an average day creating a sizable collection of images even within a short period of time, e.g. over 20,000 images per week which equates to approximately one million images captured per year. Over a lifetime of wearing this passively capturing camera, an individual could reasonably expect to have an image collection of over 50 million images. Therefore no one individual could ever manually retrieve images of encountered activities from their lives with satisfactory success. This raises the issue of how to reconstruct large personal image collections into manageable segments that can be easily retrieved by users, and to perform this segmentation automatically.

We foresee an interface whereby a user can view for each day, a number of keyframe images, each representing a different activity or event. To determine these activity representative images it is therefore imperative to automatically identify the boundaries between different activities, e.g. the ability to identify a boundary between activities such as having breakfast, working in front of a computer, having lunch with work colleagues, travelling on a bus, attending a game of football, etc.

We use 5 different sources of information to segment the SenseCam images into distinct activities, namely: low-level image descriptors, audio, temperature, light, and movement data. We will discuss how each of these 5 sources of information are processed and fused together. We will then investigate what is the minimal and best combination of data sources to carry out reliable activity segmentation based on these 5 sources.

This paper is organised as follows: The next section describes current work in this field. Thereafter the techniques used to detect activity boundaries for each data source are described in detail. We then discuss our experimental procedure followed by an analysis of the results from our experiments. Finally we determine what combination of data sources performs best followed by detailing work to be carried out in the future.

Literature Review

Several research groups have recorded personal images or audio, however their devices generally require the user to wear a laptop carried on a bag around their backs (Tancharoen, Yamasaki & Aizawa, 2006; Lin & Hauptmann, 2006), and in some cases a head mounted camera (Tano *et. al.*, 2006). As McAtamney & Parker (2006) note in their study, both the wearer and the subject talking to them are aware of personal recording devices while holding conversations. Therefore it is desirable to decrease the obviousness of the visual appearance of

a wearable device to encourage more natural interactions with the wearer. The SenseCam is small and light and from experience of wearing the device, after a short period of time it becomes virtually unnoticed to the wearer.

Gemmell *et. al.* (2004) describe the SenseCam in detail highlighting its passively capturing nature. They explain that "*…The next version of SenseCam will include audio capture, and will trigger image capture based on audio events. Eventually we would like to record audio clips surrounding image capture events…*" This motivated us to also record audio with our iRiver MP3 voice recorder. In prior work we note that audio can be a rich form of additional information that can compliment visual sources of information (Ellis & Lee, 2004a).

To our knowledge no groups have captured data for the duration of an entire day. Using the Deja View Camwear (Reich, Goldberg & Hudek, 2004) Wang *et. al.* (2006) in their work state that "*… One of the authors carried the camwear, and recorded on average of 1 hour of video every day from May to June…*" Similarly Lin and Hauptmann (2006) record data for only between 2 and 6 hours on weekdays, while others only capture for small time periods in the day too (Tancharoen, Yamasaki & Aizawa, 2005). For this paper one of the authors captured SenseCam image data for over 15 hours per day from morning to evening. This provides a more thorough representation of an individual's whole lifestyle.

One method to review images captured by the SenseCam is to use the SenseCam Image Viewer (Hodges *et. al.*, 2006). In essence this contains "*…a window in which images are displayed, and a simple VCR-type control which allows an image sequence to be played slowly (around 2 images/second), quickly (around 10 images/second), re-wound and paused…*" However it takes upwards on 2 minutes to quickly play through a day's worth of SenseCam images, which translates to 15 minutes to review all the images from 1 week. We believe a one page visual summary of a day containing images representing encountered activities or events, coupled with the ability to search for events or similar events, us a much more useful way to manage SenseCam images. Lin and Hauptmann (2006) clearly state that "*…continuous video need to be segmented into manageable units…*" A similar approach is required with respect to a lifelog collection of recorded personal images or video. Wang *et. al.* (2006) segment their video into 5 minute clips, however activities can vary in length and more intelligent techniques are required. Tancharoen and Aizawa (2004) describe a conversation detection approach in their paper. Our work is heavily focused on investigating what individual and combined data sources yield the richest activity segmentation information.

Tancharoen *et. al.* (2004) describe the benefits of recording various sources of personal information including: video, audio, location, and physiological. However they do not evaluate various combinations of these sources of data. However Wang *et. al*. (2006) investigate combining visual and audio sources to improve access to personal data and show the potential gains in using multi-modal techniques in this domain. As mentioned, they only use 2 sources of data, however in this paper we will investigate 5 sources of data and determine the optimal combinations of these sources.

Segmentation of Data into Events

The aim of automatic event detection is to determine boundaries that signify a transition between different activities of the wearer. For example if the wearer was working in front of his computer and then goes to a meeting, it is desirable to automatically detect the boundary between the segment of images of him working at the computer, and the segment of images of him being at a meeting as shown in Figure 2:

Figure 2 Example of activity boundary

We now discuss the techniques used on the various sources of data to segment each day into meaningful activities. After discussing techniques for each individual data source, we will describe our method of fusing the data sources.

Pre-Processing of Raw Data

Images were taken from the SenseCam and placed into distinct folders for each day. Use was made of the aceToolbox (AceMedia Project, 2006), a content-based analysis toolkit based on the MPEG-7 eXperimental Model (XM) (Manjunath, Salembier, & Sikora, 2002), to extract low-level image features for each and every image. The audio files recorded in parallel to the SenseCam images did not contain timestamp information and therefore it was necessary to manually note the start time of all audio segments. While this was quite tedious it will cease to be an issue with a more integrated audio functionality on the SenseCam (Gemmell et. al., 2004).

Segmentation using SenseCam Images

To process SenseCam images we used the edge histogram low-level feature, which captures the distribution of edges in an image using the Canny edge detection algorithm (Canny, 1986). Our initial intention was to use scalable colour as the low level feature, but we discovered that the edge histogram provided a better representation of event boundaries as it was less sensitive to lighting changes. Initially to search for event boundaries we carried out a form of video shot boundary detection (Brown et. al., 2000), whereby we compared the similarity between *adjacent* images using the Manhattan distance metric. If adjacent images are sufficiently dissimilar, based on a predetermined threshold, it is quite probable that a boundary between events has occurred. However this is not always the case. As the SenseCam is a wearable camera that passively captures images, it naturally captures from the perspective of the wearer. Therefore if one is talking to a friend but momentarily looks in the opposite direction an image may be taken by the SenseCam. However more than likely the wearer will then turn back to their friend and continue talking. If *adjacent* images only are compared the wearer looking momentarily in the opposite direction may trigger an event boundary, as both images could well be quite distinct in their visual nature. This feature is illustrated in Figure 3:

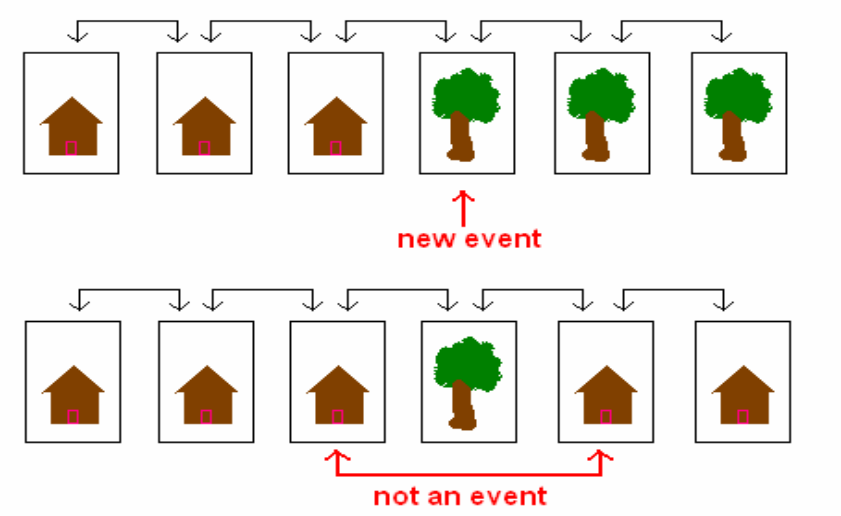

Figure 3 Illustration of possible false positive events

After communication with Gaughan & Aime (2006), we address this problem by using an adaptation of Hearst's Text Tiling algorithm (Hearst & Plaunt, 1993). This effectively involves the comparison of two adjacent blocks of images against each other, to determine how similar they are. In our work we use a block size of 5, then slide forward by 1 image and repeat the similarity calculation. If the two adjacent blocks of 5 are broadly similar then it is quite likely no event boundary has occurred, however if the two blocks are sufficiently dissimilar, based on a defined threshold after smoothing, it is quite likely that there has been a change in the wearer's activities. In using this approach the effect of outlier images, like the wearer briefly changing his point of view, is less detrimental to the detection of changes in the wearer's activities, as illustrated in Figure 4 where the house and tree icons represent two different events.

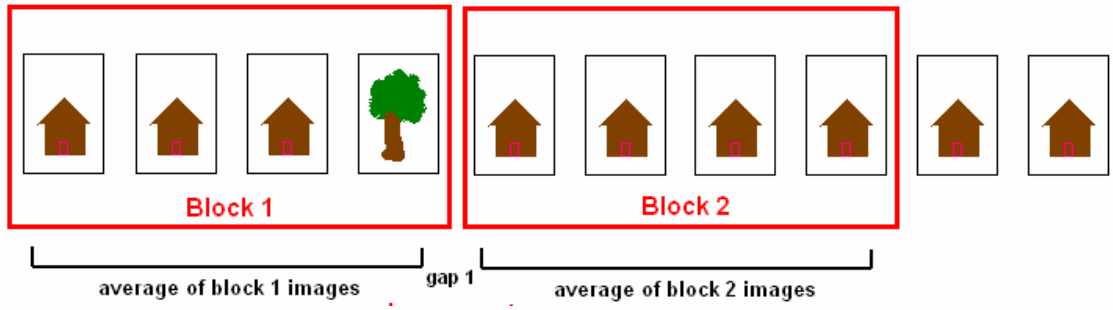

Figure 4 Image adaptation of texttiling

Segmentation using Recorded Audio

Features

Unlike speech recognition approaches which aims to distinguish audio events at a fine time scale (10 ms or 25 ms), we used long time-frame (one-minute) features that provide a more compact representation of long-duration recordings. The advantage of this is that properties of the background ambience may be better represented when short-time transient foreground events are smoothed out over a one-minute window (Ellis & Lee, 2004b). The most useful three features were log-domain mean energy measured on a Bark-scaled frequency axis (designed to match physiological and psychological measurements of the human ear), and the mean and

variance over the frame of a 'spectral entropy' measure that provided a little more detail on the structure within each of the 21 broad auditory frequency channels.

Segmentation using BIC

We used these three features to identify segment boundaries in the data using the Bayesian Information Criteria (BIC) procedure originally proposed for speaker segmentation in broadcast news speech recognition (Chen & Gopalakrishnan, 1998). BIC is a likelihood criterion penalized by model complexity as measured by the number of model parameters. Specifically, the BIC score for a boundary at time t (within an N point window) is:

$$BIC(t) = \log\left(\frac{L(X_1^N \mid M_0)}{L(X_1^t \mid M_1)L(X_{t+1}^N \mid M_2)}\right) - \frac{\lambda}{2}\Delta\#(M)\cdot\log(N)$$

where X_1^N represents the set of feature vectors over time steps 1..N etc., L(X|M)is the likelihood of data set X under model M, and Δ#(M)is the difference in number of parameters between the single model (M_0) for the whole segment and the pair of models, M_1 and M_2, describing the two segments resulting from division. The model M denotes a multivariate Gaussian distribution with mean vector μ and full covariance matrix ∑. λ is a tuning constant, theoretically one, that can be viewed as compensating for 'inefficient' use of the extra parameters in the larger model-set. When BIC(t) > 0, we place a segment boundary at time t, and then begin searching again to the right of this boundary and the search window size N is reset. If no candidate boundary t meets this criterion, the search window size is increased, and the search across all possible boundaries t is repeated. This continues until the end of the signal is reached.

In order to use the BIC approach to obtain something approximating a probability of a boundary at each time t, we must first calculate a single score for every time point. We can do this by fixing the windows used for the BIC calculation, and recording only the BIC score for the comparison of models based on two equal-sized windows either side of a candidate boundary point versus a single model straddling that point (denoted BIC_fw(t)). We can then view the BIC score as a "corrected" log likelihood ratio for that window of N points, which we can normalize to a per-point ratio by taking the Nth root. Then we can convert this to the probability whose odds ratio (p/(1-p)) is equal to that likelihood ratio i.e.

$$P(t) = \frac{\exp(BIC\_fw(t)/N)}{1+\exp(BIC\_fw(t)/N)}$$

Segmentation Using Temperature Readings

We are able to use the temperature sensor onboard the SenseCam to detect changes in location as the sensor is sensitive to within one degree Fahrenheit and thus it is possible to detect changes even when moving between rooms within the one building. To achieve this, the variance of sensor values recorded over a predetermined window size is calculated and if this is low then it is quite likely that the wearer has stayed in the same environment. However if the degree of variance is quite high it is probable that the wearer has changed environment whether by changing rooms, or perhaps going from outdoors to indoors or vice versa.

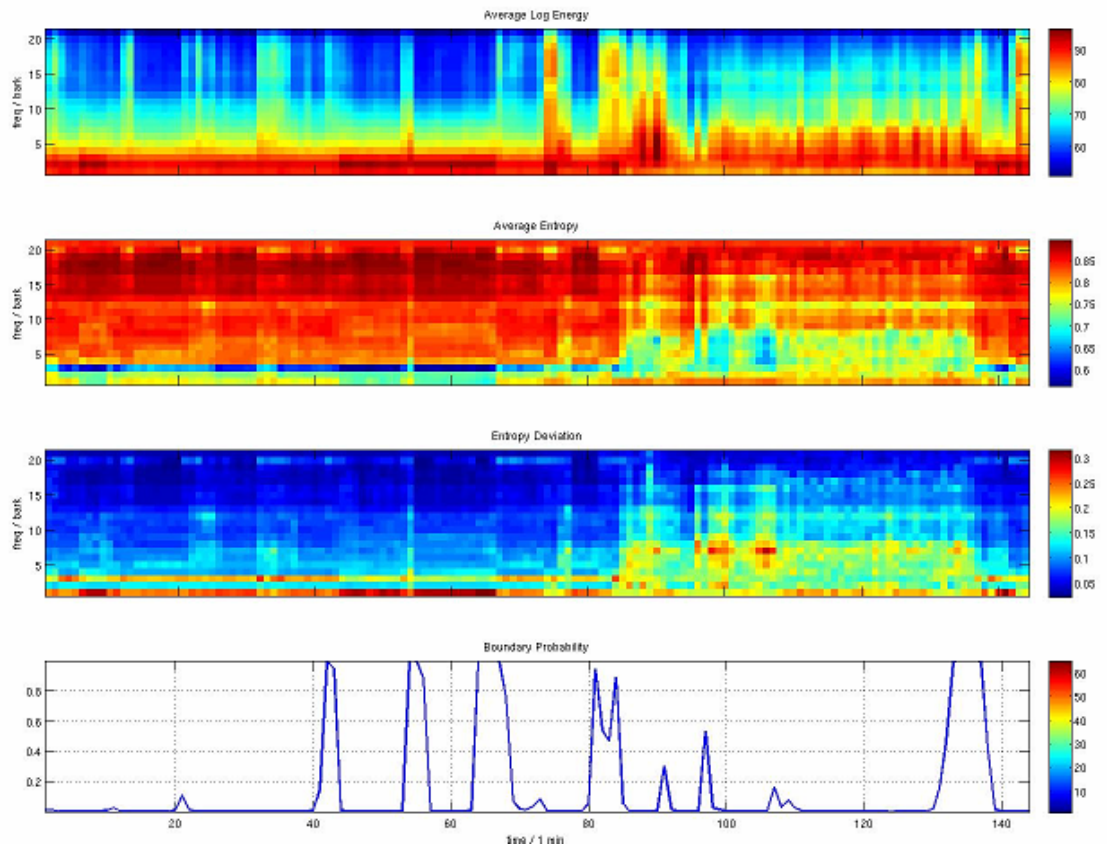

Figure 5 Plot of example audio segmentation

Segmentation Using Light Values

The light sensor onboard the SenseCam simply measures the intensity of white light present. To calculate the likelihood that a given image represents a boundary between different events we compute the derivative value of the white light sensed at the image time, i.e. how much of a change in lightening there was. If there is a small amount of change in the light level is it quite probable that the wearer has remained in the environment that they were already present in. However if there is a substantial difference in the light level sensed it is entirely possible that the wearer has in fact changed environment, e.g. if moving from one room to another, as the wearer approaches the door the light level will decrease, however once they open the door and enter their new environment the light level quickly increases again. This may indicate a change in activities by the wearer. High changes in light level would not only point towards the wearer moving to a different location as it could also indicate to something happening within the same location e.g. when the wearer talks to a friend within the same room, they may just simply have turned in a different direction in the room which may have a very different lighting level.

Segmentation Using Accelerometer Values

Motion of the SenseCam is calculated using the 3-axis accelerometer data captured by the device. We follow the approach of Ó Conaire et. al. (2007) in determining the volume of motion associated with each image. Firstly we compute the derivative for each of the X, Y, and Z accelerometer values, as we're interested in the rate of change in motion. E.g. if one is sitting down for lunch, it is quite likely that a high degree of motion would correlate to the wearer being finished lunch and walking back to their workspace. After computing the derivative value

for each axis, we combine the 3 different axes using: $\sqrt{X^2 + Y^2 + Z^2}$ Each sensor value is then smoothed using a median filter.

Finally since sensor data is captured every 2 seconds, as opposed to an average of 20-25 seconds for images (based on our data collection of 10 days), the sensor motion values are associated with an image using a Gaussian window centred at the capture time of the image (Ó Conaire *et. al.*, 2007). Larger motion values indicate that the associated image is quite likely to represent a boundary between different events or activities e.g. walking from home to work, walking from the office to lunch, walking from home to the shop, etc.

Fusion of Data Sources

As the SenseCam image and sensor data is automatically time-stamped, the challenge was that of combining the output of audio data analysis with analysis of the other data sources. Due to privacy concerns it was not possible to record audio at all times therefore there were occasions when image and sensor data was present, but no audio data was recorded. By nature the BIC segmentation dissects the recorded audio file into equal sized frames of one minute each, whereas the interval between images is variable. Thus to align the audio data with the image data, it was decided to linearly interpolate the scores from the one minute audio frames to provide a more dense format, at the scale of one second.

The segmentation confidence scores of each data source were normalised using the Sum normalisation method of Montague & Aslam (2001).

To fuse the actual sources of data we decided to use the CombMNZ approach (Fox & Shaw, 1993). This approach was chosen as it considers sources with scores of zero, a desirable property given that audio was not recorded at all times due to privacy concerns. In this approach the sum of all confidence scores is multiplied by the number of non-zero data sources to give an overall combination value for the image in question. This fusion metric, where each contributor is weighted equally, is calculated for each image in each day and is thus a form of early fusion.

For each data source, and each fusion combination of data sources, a fixed number of activity segmentations were chosen as the output; in our case, for each day, the 20 top-scoring images were considered most likely to be activity boundaries. No two images occurring within 5 minutes of each other were allowed due to the naturally broad Gaussian distribution of scores around a possible boundary. All possible combinations of the 5 data sources were computed to give 31 different sets of boundaries.

Experimental Set-up

The definition of an activity is largely subjective and also has many different levels of granularity. For example one person may describe a normal day as being at home, driving to work, being at work, driving home, talking to family, going to bed. However within each of these events are many sub-events such as work being broken down into being at a computer, at lunch, at a computer, important meeting, and back to the computer. Again within each of these it is possible to further breakdown the task, and so on. As a result of this it is difficult to determine a true ground truth of segmentations on personal data. Therefore, although it is feasible to do so, we feel that there is an inherent flaw in asking the wearer of a SenseCam to

manually create a ground truth of segmentations for all his/her SemseCam images and then to compare automatically generated results against this.

To evaluate the validity of one segmentation approach against other segmentation approaches we instead will examine the segmentations of an approach which are unique in comparison to the segmentations of other approaches. We assume it is not necessary to comment on boundaries common to all the approaches in question, therefore we just manually examine the results unique to each approach. Unique segments are determined as those detected by one data source, with no segments detected over a 15 minute window by the other data sources.

One of the authors collected SenseCam image and sensor data along with audio data from the iRiver MP3 audio recorder over a 10-day period between late August and early September of 2006. Image and sensor data was collected for the entire duration of each day, from morning to night. Due to privacy concerns audio data could not be recorded in all locations at all times therefore only approximately 80 hours of audio data was collected during this period of time. Data logging started on a Friday, a day that included a meeting with work colleagues. Then for the next two days the author was at home visiting his family and friends and attended a football match, before getting on a bus back to his residence near work. The next four days all involved normal routine work activities, along with talking to his housemate in the evening times. The last day of recording included a flight abroad and involved normal activities experienced in an airport. Figure 6 below illustrates what these activities look like.

Figure 6 Example of various activities

Presentation of Results

The results presented below were calculated on an image set of 22,173 images over the aforementioned 10 day period. Every evaluated sensor, and combination of fused sensors, produced 200 activity boundaries for this dataset (20 per day as described in fusion of data sources section).

We now investigate the relative merits of each individual data source measured against the other data sources, for unique detection of events. Table 1 below details the precision results for uniquely identified segmentations (compared to the other independent data sources) for each of the five individual sources of data over the collection of 10 days.

Data Source	Precision
Image	21/46
Audio	7/44
Temperature	9/27
White Light	5/17
Accelerometer	6/16

Table 1 Precision of valid unique segmentations

Segmentation of SenseCam Images Using Image Features

Over the collection of data there were 46 unique activity boundary segmentations as detected by the image processing technique. Of those, 21 segmentations were deemed to be true segmentations of which 11 represented boundaries between activities that involved the wearer moving to a different location, e.g. walking from the wearer's house to their car, wearer getting off plane at airport, etc. The other 10 true segmentations represented boundaries between activities that occurred in the same location, e.g. the wearer talking to a colleague in work (Figure 7), or talking on the phone in the wearer's living room, etc.

Figure 7 Boundary detection from working at computer to talking to colleague

26 segments were falsely detected as boundary changes between wearer activities. The most common reason to trigger these false events appeared to be due to the wearer slightly changing their seating position while working in front of the computer, or changing their standing position while talking to friends. 10 such events were detected (Figure 8). A plausible reason for these detections is that the block of images prior to the change will be significantly different due to the images being taken from a slightly lower or higher seating/standing position. This appears to sufficiently change the edge properties of the images. 6 segmentations were falsely detected due to movement by the wearer of the SenseCam e.g. walking around within the same room, going to the fridge while in the kitchen, etc. There were 4 invalid events detected due to the wearer briefly holding some object in front of them, e.g. searching for an item in a bag. If the bag is held close enough to the camera that one or two particular images are greatly different from the normal environment images, an activity change can be triggered. Finally lighting changes in buses or cars at nighttime significantly change the edge properties of the images so that an activity boundary is falsely detected. This occurred 5 times.

Figure 8 Examples of false positive boundary due to adjusting seating position

Segmentation of SenseCam Images Using Audio Features

Over the collection of data there were 44 unique activity boundary segmentations detected by the audio processing technique and of those only 7 were deemed true activity boundaries. Of those 7 positive segmentations, 3 were due to the wearer talking to another person (Figure 9). The other 4 segmentations were due to background noise and wearer-generated sounds e.g. packing items into a travel bag.

Figure 9 Boundary of talking to colleague

36 segments were falsely detected as boundary changes between activities. 30 of those events were due to significant background noise, e.g. television () and radio noise when the wearer was working in front of his computer, the noise of the plane landing/taking off while onboard an airplane, crowd noise while at a football match, etc. The 7 other falsely detected segments had no significant changes in audio activity. A possible reason for this is the fact that 20 events are chosen for each day, and the audio was recorded for a shorter duration than normal on the days of those 7 events. It is quite probable that there was not 20 events from those days during the time that the audio was recorded.

Segmentation of SenseCam Images Using Temperature Readings

There were 27 segmentations uniquely identified by the temperature sensor. Of those, 27 segmentations, 9 were determined to be valid activity boundaries. In all cases except one the segmentations were caused when the wearer changed room. This is to be expected as there is generally a change in temperature when moving from one room to another. It is interesting to note that the other positively detected activity involved the user having to go under his desk to change an electrical plug and the change of temperature in that part of the room (near the floor as opposed to normal seating/standing position) was detected (Figure 10).

Figure 10 Boundary of wearer installing new hardware device

18 segmentations were falsely detected as boundaries between activities using temperature readings. Of those, 6 were due to the actual SenseCam device being powered on. It appears to take a short period of time for this sensor to calibrate and thus a small variance in measured

temperature can occur. The other 12 falsely detected segmentations were due to a change in ambient temperature in the actual environment that the user was in. 7 such segmentations occurred indoor and the other 5 happened outdoors. Possible explanations include the indoor airconditioning being switched on or off, and for outdoor activities it is possible that temporary cloud cover affected the ambient temperature sufficiently to trigger an event.

Segmentation of SenseCam Images Using White Light Levels

There were 17 segmentations uniquely identified by the white light sensor. Of those, only 5 were identified as valid boundaries between user activities. It is interesting to note that 4 of those activities occurred within the same room, e.g. the wearer talking to a work colleague in a lab environment (Figure 11). The one other event required the wearer to exit the lab environment to answer a phone call.

Figure 11 Boundary of talking to work colleague

12 segmentations were falsely detected as boundaries between activities. As with the image source of data, the white light data source is suspectible to objects being held closely to the actual SenseCam device (Figure 12). This naturally effects the level of light detected, sufficiently so much that an activity boundary is detected. Of the 12 false segmentations, 8 were due to sunlighting changes, predominantly when the wearer was in a car or bus. The lighting changes were magnified due to rays of sunlight coming in through the windscreens of the car or bus.

Figure 12 False positive boundary when wearer moving paper in bus

Segmentation of SenseCam Images Using Accelerometer Values

There were 16 segmentations uniquely identified by the accelerometer sensor. Of those, 6 were identified as valid boundaries of user activities. From those 6 positive segmentations it is interesting to note that 5 of them involved the user moving to a different room e.g. Figure 13. The one other positive segmentation involved the wearer sitting down in an airport while waiting for a flight.

Figure 13 Boundary of wearer walking to different location

10 segmentations were falsely detected as boundaries between activities. Significantly all 10 of these events involved wearer movement, however they were not judged as sufficiently important to signify a change in user activities. Examples include the wearer moving around his

seat on a bus to find a travelling bag at his feet (Figure 14), reaching to the floor to take a drink of bottled water while working in front of a computer, walking around an airport, etc.

Figure 14 False positive boundary of wearer moving in bus seat

Segmentation of SenseCam Images Using Combinations of Data Sources

After viewing each data source autonomously it is interesting to note that some have similar strengths. 3 sources of information regularly detected the wearer changing location, namely image, temperature, and accelerometer whereas only half of the valid image processed boundaries were due to location change, almost all the boundaries derived by the temperature and accelerometer sensors were due to location change. Therefore we investigated the relative merits of the temperature and accelerometer sensors to see which yielded more rich location change results. These will be compared against each other and also against a fusion of both sources. The aim here is to find out if we can use just one of those two sources to provide good change of location information, or whether it will be necessary to fuse both together to derive this information in a sufficiently accurate manner

Similarly the light and image sensors were the optimal sources of information to detect events occurring within the same location. We investigated the merits of each of these sensors in determining if we need to combine them or not in order to provide the most rich segmentation results.

The best segmentations provided by the audio sensor were times of social interaction. The image and light sensors are best at picking up boundaries within a static environment, therefore we also investigated the merits of the fusion of these two sensors against the audio data source in determining boundaries containing social interaction.

Optimum Location Change Segmentation

The temperature and accelerometer sensors were fused together using the CombMNZ approach as described earlier. The output of the temperature sensor, the accelerometer sensor, and the fusion of both were compared against each other to find the unique segments that each produced. Judgments were then made on the unique segmentations produced by each of the 3 approaches. This approach is applied to all subsequent fusions. Table 2 displays the precision of valid location changes identified among the unique segmentations of each approach.

Data Source	Precision
Temperature	17/48
Accelerometer	17/45
Fusion of both	0/3

Table 2 Uniquely identified location change segmentations

From observation of the valid segmentation boundaries it is interesting to note that the accelerometer performs better at identifying room changes within the one building (15 of the 17 segmentations), whereas the temperature sensor is stronger in detecting when the user moves

from outdoors to indoors and vice-versa (7 such unique detections as against 2 by the accelerometer). This is to be expected as the variance in temperature values will be greater moving between outdoor and indoor environments than moving between rooms in a building. It is also interesting to note that the fusion of both sources does not provide any positive uniquely identified location change boundaries. We thus feel, based on our observations, that the accelerometer sensor alone is the best source of information in determining location changes. There is little difference between the temperature and accelerometer sources of information, however from viewing the results in Table 1 it is clear that the accelerometer sensor provides less false positive events on its own.

Optimum Activity Boundaries in the Same Location

Here the image processing and light sensors were compared against each other and against the fused results of both. Table 3 displays the precision of valid within location changes identified among the unique segmentations of each source.

Data Source	Precision
Image processing	26/81
Light sensor	10/37
Fusion of both	4/7

Table 3 Uniquely identified within location segmentations

It is immediately striking that while the image processing technique does provide the most positive detections of activities, it also produces a large number of false positive boundaries. From observation of the detected boundaries it is noticeable that the image processing technique is better at detecting more fine-grained events e.g. beginning to work on laptop in airplane (Figure 15). However we do feel that the fusion of both sources represents the best solution as there are fewer false positive segmentations provided than by the image or light sensor sources used autonomously.

Figure 15 User beginning to work on laptop

Optimum Social Interaction Segmentation

Here we compared the audio processing to the fusion of image and light sensors to investigate which approach yields the optimal segmentations induced by social interaction. Table 4 displays the precision of valid boundaries triggered by social interaction among the uniquely identified segmentations of each source.

Data Source	Precision
Audio	18/51
Image/Light fusion	20/60
Fusion of audio/image/light	1/12

Table 4 Uniquely identified social segmentations

The audio source of information is very good at detecting events of the wearer talking to other people in scenes that are almost static in a visual sense (13 of the 18 uniquely detected boundaries), e.g. Figure 16 where the wearer was talking to the passenger beside him in the

airplane. On the other hand the fusion of image and light sources is naturally dependent on scene changes. E.g. Figure 11 where the scene changes from the monitor being dominant in the image to a different desk and a person becoming dominant in the image. While it is evident from the results that the uniquely identified segmentations by the fusion of all three sources does not yield positive segmentations, it is difficult to ascertain which of the other two approaches is optimal in detecting social interaction induced boundaries.

Figure 16 Talking to passenger seated beside wearer (note hand gestures)

Optimal Segmentation Results

In the final set of experiments we compared three distinct fusions of the data sources to investigate which one provides the richest segmentation results. We believe that superior combinations should provide a mix of boundaries induced by the wearer changing location, by the wearer starting a different activity within the same location, and by the wearer socially interacting with others. Based on our observations heretofore we propose the following combinations of sensors for event detection:

Audio and accelerometer sources: The reasoning for this is that the audio would detect social interaction boundaries and also some boundaries within the same location too, while the accelerometer would detect changes both between different locations, and also within the same location.

Image, light, and accelerometer sources: The image and light sensors are intended to detect both changes within the same location as well as boundaries triggered due to social interaction. The accelerometer would detect both boundaries between locations and also within locations.

Image, audio, temperature, light, and accelerometer sources: Here all data sources are used and it could be reasonably expected that all possible boundary types would be detected.

Data Source	Precision
Audio/Accelerometer	12/42
Image/Light/Accelerometer	39/61
All data sources	7/14

Table 5 Uniquely identified segmentations

We computed event detections for these 3 distinct combinations of sensors and Table 5 displays the precision of valid boundaries among the uniquely identified segmentations from each combination. It is interesting to note that of the 12 audio/accelerometer valid segmentations, 5 were social interaction events, 5 were activity boundaries within the same location, but only 2 were changes in location. This would indicate that perhaps the accelerometer sensor did not have as large an influence on the fused results as it did in the other 2 combinations. It is indeed encouraging that a large proportion of the valid boundaries were social interaction activities.

The combination of the image, light, and accelerometer data performed best of the three evaluated combinations. Of the 39 valid segmentations, 21 were due to the wearer changing location. This indicates that the accelerometer data was a significant, but positive, influence on the fused results. The other 18 valid segmentations were evenly split between boundaries in activity in the same location, and boundaries in wearer activity due to social interactions.

The combination of all data sources fused together provided much less unique segmentations than the other two combinations. Half of those segmentations were valid boundaries between different user activities. 3 of the valid segmentations were due to social interaction, while 2 were due to changes in user activity in the same location. There was 1 valid boundary due to the user changing location. This indicates that the temperature and accelerometer sensors may not have influenced the fused results to the same degree as the other three data sources.

Conclusions and Future Work

Initially we investigated unique event segmentations identified by each autonomous data source in comparison to the other autonomous data sources. From this we were then able to identify three main types of activity boundaries, namely: a change of activities within the same location, the wearer moving to a different location, and the wearer being engaged in some form of social interaction. For activity segmentation within the same location we found that a fusion of image processing and light sensor processing performed best. We believe that the accelerometer sensor alone provides better segmentation of activities where the wearer moves to a different location. Finally we believe that either the audio sensor alone or the fusion of the image and light processing sources provides better detection of boundaries induced by the wearer beginning to socially interact with others, than the fusion of all three aforementioned sources.

From our experience, the optimal combination of sensors for all types of event detection proved to be a fusion of the image, light, and accelerometer data sources. This combination provided quite a large number of additional valid segmentations than two other combinations that we tried. The fusion of all 5 data sources produced the least number of false positives however we still feel that it did not provide a sufficient number of positive segmentations to prove more valuable than the combination of image, light, and accelerometer data. An additional advantage of discarding 2 of the data sources is the reduced processing load.

In future we feel that it would be prudent to investigate the benefit of other data sources towards activity segmentation. Previously we have segmented GPS data into different trips (Doherty *et. al.*, 2006) and believe that a refinement of this may be helpful in identifying activity changes by the wearer on a larger scale to compliment the activity change that the accelerometer is proficient at identifying at the room level.

In addition to investigating other sources of data, it will be important to investigate more closely the optimal method of normalising the data sources to be put forward for data fusion. Other methods of fusion should also be compared to see which performs best on a lifelog of personal images.

In this paper we have argued that segmenting SenseCam images into meaningful activities is very important if an archive of SenseCam images is to be useful as a memory augmentation aid or indeed in any other application for SenseCam images. We also feel that it will be important to determine which of those activities is the most important in order to start distinguishing

events by something other than time of occurrence. This is desirable as more important activities can be given greater emphasis by an interface helping the user to review images summarising their activities.

Acknowledgements

The SenseCam project at Dublin City University is supported by Microsoft Research. We are grateful to the AceMedia project for use of the aceToolbox toolkit. We would like to thank the Irish Research Council for Science, Engineering and Technology; Science Foundation Ireland under grant number 03/IN.3/I361 for support; and also the European Commission under contract FP6-027026-K-SPACE. This paper reflects the views of the authors only and not necessarily by any of the aforementioned partners. Finally we thank the reviewers for their informative feedback.

References

Aime, S., & Gaughan, G. (2006). Personal communication.

The AceMedia project. *http://www.acemedia.org*, Last accessed 30 November 2006.

Brown, P., Smeaton, A.F., Murphy, N., O'Connor, N., Marlow, S., & Berrut, C. (2000). Evaluating and Combining Digital Video Shot Boundary Detection Algorithms. *IMVIP 2000 – Irish Machine Vision and Image Processing Conference.*Belfast, Ireland.

Canny, J. (1986). A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 8(6).

Chen, S., & Gopalakrishnan, P. (1998). Speaker, environment, and channel change detection and clustering via the Bayesian Information Criterion. *In Proceedings DARPA Broadcast News Transcription and Understanding Workshop.*Lansdown, Virginia, USA.

Doherty, A.R., Gurrin, G., Jones, J.F., & Smeaton, A.F. (2006). Retrieval of Similar Travel Routes Using GPS Tracklog Place Names. *SIGIR GIR – Conference on Research and Development on Information Retrieval, Workshop on Geographic Information Retrieval.* Seattle, Washington, USA.

Ellis, D.P.W., & Lee, K. (2004a). Minimal-Impact Audio-Based Personal Archives. *CARPE 04 - First ACM workshop on Continuous Archiving and Recording of Personal Experiences.* New York, USA.

Ellis, D.P.W., & Lee, K. (2004b). Features for segmenting and classifying long-duration recordings of personal audio. *In Proceedings ISCA Tutorial and Research Workshop on Statistical and Perceptual Audio Processing SAPA-04*. Jeju, Korea.

Fox, E., & Shaw, J. (1993). Combination of Multiple Searches. *TREC 2 – Text REtrieval Conference. Gaithersberg, Maryland, USA.*

Gemmell, J., Williams, L., Wood, K., Lueder, R., & Bell, G. (2004). Passive Capture and Ensuing Issues for a Personal Lifetime Store. *CARPE 04 - First ACM workshop on Continuous Archiving and Recording of Personal Experiences.* New York, USA.

Hearst, M.A., & Plaunt, C. (1993). Subtopic structuring for full-length document access. *SIGIR - Proceedings of the 16th Annuan ACM-SIGIR Conference on Research and Development in Information Retrieval.*Pittsburg, USA.

Hodges, S., Williams, L., Berry E., Izadi, S., Srinivasan, J., Butler, A., Smyth, G., Kapur, N., & Wood, K. (2006). SenseCam : A Retrospective Memory Aid. *UbiComp – 8th International Conference of Ubiquitous Computing.* California, USA.

Lin, W., & Hauptmann, A. (2006). Structuring Continuous Video Recordings of Everyday Life Using Time-Constrained Clustering. *Multimedia Content Analysis, Management, and Retieval – In Proceedings of SPIE-IS&T Electronic Imaging*. San Jose, California, USA.

Manjunath, B., Salembier, P., & Sikora, T. (2002) Introduction to MPEG-7 : Multimedia Content Description Language. *John Wiley & Sons.*

McAtamney, G., & Parker, C. (2006). An Examination of the Effects of a Wearable Display on Informal Face-To-Face Communication. *SIGCHI – Proceedings of the SIGCHI conference on Human Factors in computing systems*. Montréal, Québec, Canada.

Montague, M., & Aslam, J. (2001). Relevance Score Normalization for Metasearch. *CIKM – Proceedings of the tenth international conference on Information and knowledgement management*. Atlanta, Georgia.

Ó Conaire, C., O'Connor, N.E., Smeaton, A.F., & Jones, G.J.F. (2007). Organising a Daily Visual Diary Using Multi-Feature Clustering. *SPIE Electronic Imaging – Multimedia Content Access : Algorithms and Systems.*. San Jose, California, USA.

Reich, S., Goldberg, L., & Hudek, S. (2004). Deja View Camwear Model 100. *CARPE 04 - First ACM workshop on Continuous Archiving and Recording of Personal Experiences.* New York, USA.

Tancharoen, T., & Aizawa, K. (2004). Novel Concept for Video Retrieval in Life Log Application. *PCM –Pacific Rim Conference on Multimedia.* Tokyo, Japan

Tancharoen, D., Yamasaki, T., & Aizawa, K. (2005). Practical Experience Recording and Indexing of Life Log Video. *CARPE – Second ACM workshop on Capture, Archival and Retrieval of Personal Experiences.* Singapore.

Tancharoen, D., Yamasaki, T., & Aizawa, K. (2006). Practical Log Video Indexing Based on Content and Context. *Multimedia Content Analysis, Management, and Retieval – In Proceedings of SPIE-IS&T Electronic Imaging*. San Jose, California, USA.

Tano, S., Takayama, T., Iwata, M., & Hashiyama, T. (2006). Multimedia Informal Communication by Wearable Computer based on Real-World Context and Graffiti. *ICME – IEEE International Conference on Multimedia & Expo.* Toronto, Ontario, Canada.

Wang, Z., Hoffman, M.D., Cook, P.R., & Li, K. (2006). Vferret : Content-Based Similarity Search Tool for Continuous Archived Video. *CARPE – Third ACM workshop on Capture, Archival and Retrieval of Personal Experiences.* Santa Barbara, California, USA.