Providing Effective Memory Retrieval Cues through Automatic Structuring and Augmentation of a Lifelog of Images

Aiden R. Doherty B.Sc. (Hons)

A Dissertation submitted in fulfilment of the requirements for the award of

Doctor of Philosophy (Ph.D.)

to the

Dublin City University

School of Computing,

Centre for Digital Video Processing (CDVP)
& CLARITY: Centre for Sensor Web Technologies

Supervisor: Prof. Alan F. Smeaton

December, 2008

This thesis is based on the candidate's own work, and has not previously been submitted for a degree at any academic institution.

Declaration

I hereby certify that this material, which I now submit for assessment on the programme of study leading to the award of Doctor of Philosophy is entirely my own work, that I have exercised reasonable care to ensure that the work is original, and does not to the best of my knowledge breach any law of copyright, and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the text of my work.

Aiden R. Doherty (ID: 55146261)

Abstract

Lifelogging is an area of research which is concerned with the capture of many aspects of an individual's life digitally, and within this rapidly emerging field is the significant challenge of managing images passively captured by an individual of their daily life. Possible applications vary from helping those with neurodegenerative conditions recall events from memory, to the maintenance and augmentation of extensive image collections of a tourist's trips. However, a large lifelog of images can quickly amass, with an average of 700,000 images captured each year, using a device such as the SenseCam.

We address the problem of managing this vast collection of personal images by investigating automatic techniques that:

1. Identify distinct events within a full day of lifelog images (which typically consists of 2,000 images) *e.g. breakfast, working on PC, meeting, etc.*

2. Find similar events to a given event in a person's lifelog *e.g. "show me other events where I was in the park"*

3. Determine those events that are more important or unusual to the user and also select a relevant keyframe image for visual display of an event *e.g. a "meeting" is more interesting to review than "working on PC"*

4. Augment the images from a wearable camera with higher quality images from external "Web 2.0" sources *e.g. find me pictures taken by others of the U2 concert in Croke Park*

In this dissertation we discuss novel techniques to realise each of these facets and how effective they are. The significance of this work is not only of benefit to the lifelogging community, but also to cognitive psychology researchers studying the potential benefits of lifelogging devices to those with neurodegenerative diseases.

Acknowledgements

Doing a PhD is a wonderful experience and one that I have thoroughly enjoyed. I am very fortunate to be in the position to complete this PhD and it is entirely because of the support I have had around me. The most important people in my life are obviously my family and to be honest they have made life exceptionally easy and brilliant for me. Thanks to Denise and Maria for being great sisters, and especially to Mum (Teresa) and Dad (Denis) for creating all the amazing opportunities I have enjoyed in my life thus far.

Outside of my family the most influential person in my life has been my supervisor, Alan F. Smeaton. I would sincerely like to thank him for firstly taking the risk in supervising me for my PhD. He has always been very supportive of any ideas or initiatives I have been interested in pursuing, and he has also been very proactive in allowing me attend many conferences and partake in numerous research visits. Not only has this benefited me professionally, but perhaps as importantly it has matured me as a person and helped me become more international in my thoughts. I simply can't thank Alan enough.

A major reason that my PhD has been such a pleasurable experience is due to the fact that I was sponsored by the Irish government through the Irish Research Council for Science, Engineering and Technology (IRCSET). As a result of this fantastic sponsorship I never had any financial worries throughout my PhD, and also I have been able to travel extensively to further my research. Their support also allowed me to purchase a laptop which has been vital towards my research. I also wish to thank Science Foundation Ireland (SFI) who has supported my research through their sponsorship of the Adaptive Information Cluster (AIC) and CLARITY. Being part of these SFI funded research groups has given me endless research opportunities through the extensive resources available. Additionally I wish to extend my sincere gratitude to Microsoft Ireland for their sponsorship in the final year of my PhD. This support has allowed me to purchase extra equipment and also travel to a conference which has further developed my skill set as a researcher. I would also like to thank Microsoft Research in Cambridge, U.K., for providing CLARITY with SenseCam devices, as otherwise my PhD topic would not have been possible. I would like to thank Dublin City University for their accommodation scholarship in the first year of my PhD studies. It is always tough moving somewhere new, but a fantastic living environment made my time here in Dublin really enjoyable.

I would like to thank my external and internal examiners for their valuable feedback, Steve Hodges and Andrew Salway. Throughout my PhD I have had the opportunity to visit many other research institutions, and I am indebted to the following people for their openness, friendliness, hospitality, assistance, and time: Jim Gemmell (Microsoft Research Redmond); Dan Ellis (Columbia University); Steve Hodges and James Scott (Microsoft Research Cambridge); Alex Hauptmann (Carnegie Mellon University); Paul McKevitt and Kevin Curran (University of Ulster); Kalervo Järvelin (University of Tampere, Finland); Chris Moulin (University of Leeds); and Roelof van Zwol (Yahoo! Research Barcelona). These visits were all very beneficial towards my research and I am most grateful to all these wonderful people.

Given that I have spent my last 3 years in the Centre for Digital Video Processing (CDVP), I simply can't speak highly enough of everyone working here. The level of knowledge and expertise is staggering, but most importantly the people working here have fantastic personalities. Everyone in the group helped at various stages throughout my time here in the CDVP.

Lastly it has been a whirlwind three years full of many highs and natural lows. In experiencing the lows of death and near-death experiences, it teaches you to appreciate the time spent with friends. I'd like to thank people in and around work, especially the regular lunch crew, all the folk over in the gym, my house mates and importantly all my friends from back home who I could only meet every so often either down here in Dublin or back home up in Donegal.

Contents

List of Figures

List of Tables

Chapter 1

INTRODUCTION

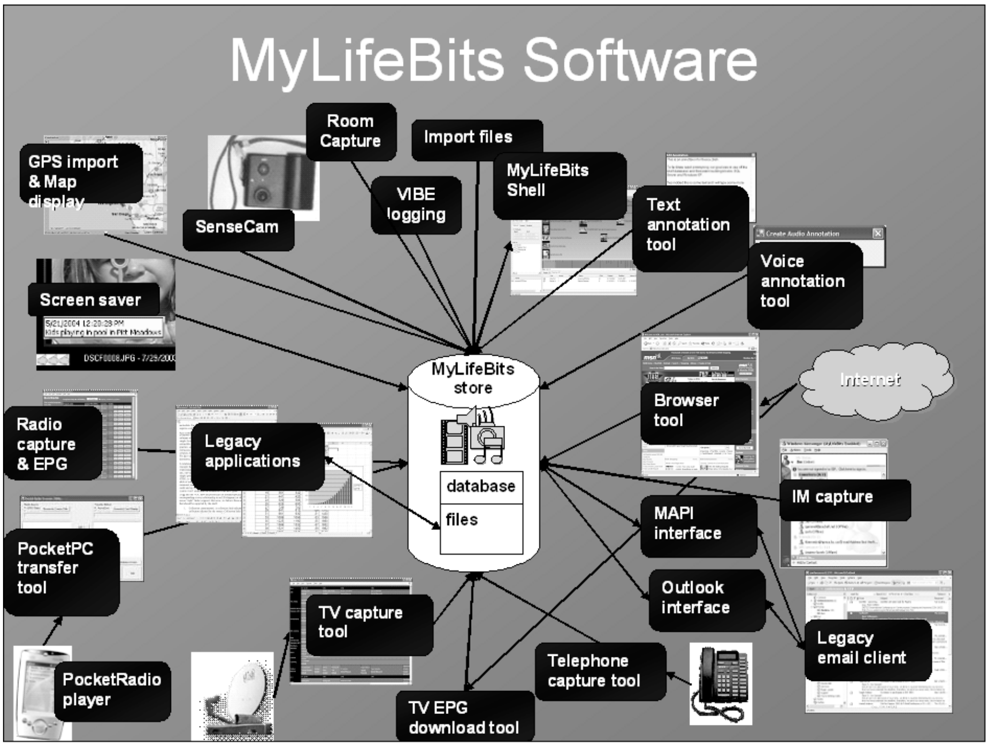
The important events in our lives define our being, and it is important to remember these events. In fact Conway argues that autobiographical memory affects us to the very core and defines who we are [40]. McAdams also argues that memory is crucial to self-definition [98], and Lampinen, Beike, & Behrend acknowledge the contribution of Sir Fredric Bartlett who recognised in 1932 that memory was not a mechanical process but a meaning-making system [84]. In the past exaggeration and repeated storytelling were used to remember. In fact shared memories are a part of our social self, and define who we are as a group of people e.g. it is the memories of shared experiences with one's family that create the close ties and bonds, and reminiscing these memories together not only improves one's memory, and the memory of the family, but also create a tighter social bond between those people. Therefore in essence memories do affect us to the core, and in fact define our relationships too. Given that populations are generally getting older and that the number of those with memory impairments is ever increasing [9], information scientists are now exploring using technology to aid memory.

Almost everything we do these days is in some way monitored or logged by computers. We've come to accept, or maybe just ignore, this massive surveillance because it brings us benefits. We have a more secure feeling when we know there is CCTV present, we get itemised billing from phone companies, and we get convenience and even loyalty bonuses with some of our regular purchases.

Lifelogging is the term used to describe recording different aspects of our daily life, in digital form, for our own exclusive personal use. It is a form of reverse surveillance, sometimes termed *sous*veillance, referring to us, the subjects, doing the watching of ourselves. Lifelogging can take many forms, such as the application which runs on a mobile phone to 'log' all your activities and then present all those activities in a calendar format.

Human memory is recognised as being far from perfect and through the ages human beings have devised many methods to aid their memory [9], from ancient cave paintings, to verbal story telling with embellishments, to written manuscripts to maintain a more accurate historical account, to audio and video recorders in the recent past. In terms of individuals remembering aspects/activities of what they themselves have been doing, the written diary has been popular for a long number of years. More recently, with the advent of computing technology and the Internet, many individuals have begun to maintain online blogs to detail aspects of their activities.

One of the primary goals of *lifelogging* is to move individuals towards total memory/experience recall [15]. The lifelogging community attempt to achieve this goal by automatically capturing electronic data from numerous sources of information (Figure 1.1), e.g. web pages visited, e-mails sent and received, audio recordings of conversations, etc. An interesting area in this field of research is visual lifelogging, i.e. an individual capturing activities through the medium of images or video. It is particularly important to do this through passive capture, meaning the use of devices that automatically capture images or video, thus requiring no conscious effort by the user to take images, which leads to him or her acting in a more natural manner. Baddeley recognises the use of cave-paintings and visual mnemonics (first known to be used by the Greek poet Simonides in about 500 B.C.) through the ages to aid the remembrance of past events [9], thus highlighting the importance of visual imagery to memory (as found by Brewer in 1986 [24]). We will later show that cue-based recall is very helpful in neural retrieval mechanisms, in particular when those cues are lifelog images taken from one's own perspective.



MyLifeBits Software

Room Capture

Import files

GPS import & Map display

SenseCam

VIBE logging

MyLifeBits Shell

Text annotation tool

Voice annotation tool

Screen saver

MyLifeBits store

database

files

Internet

Browser tool

Radio capture & EPG

Legacy applications

IM capture

MAPI interface

PocketPC transfer tool

Outlook interface

TV capture tool

Legacy email client

PocketRadio player

Telephone capture tool

TV EPG download tool

Figure 1.1: Overview of MyLifeBits, a project exploring the challenges posed in recording and retrieving personal information from multiple sources of data [60].

The visual lifelogging community has mainly concentrated on the challenges of miniaturising the hardware devices, and of how to store the *vast* quantities of data. Only recently has the focus been placed on how to manage and organise these large quantities of data to act as effective memory retrieval cues.

We will now firstly provide more details on the SenseCam, a passive capture visual lifelogging device. Thereafter we will detail potential scenarios where such a device may be of real-life benefit to individuals, whether they be tourists or have neurodegenerative conditions. While capture and storage of visual lifelog images is relatively straightforward, a major challenge which has faced the community is that of effectively managing and retrieving relevant content. We will discuss in detail the particular challenges posed by lifelogging, and detail a number of focused areas in which we make a contribution. Finally in this chapter we state the hypothesis of this thesis and provide a structure of how the remainder of this thesis is organised.

1.1 The SenseCam

We will now describe the visual lifelogging device that we use, the SenseCam, in more detail. The SenseCam [71], developed by Microsoft Research Cambridge, is a small wearable device that is worn via a lanyard around the neck as displayed in Figure 1.2. The SenseCam 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 infrared 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 [71]. 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. Images are stored onboard the device, with an average of about 2,000 images captured in a typical day, along with associated sensor readings. The SenseCam is capable

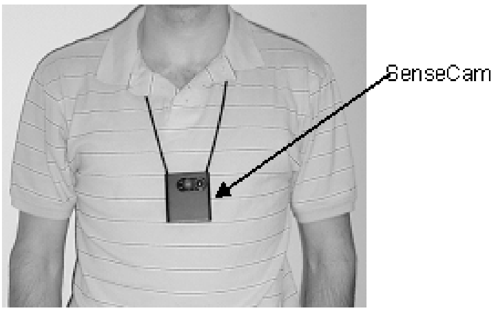


Figure 1.2: Note the SenseCam device worn via a lanyard around the user's neck.

of storing approximately 15 days worth of data onboard, before it is required to download the images to a standard PC via a USB cable. It is required to charge the SenseCam every night.

Other work in the literature has touched on potential applications based on pictures and sensor values recorded by the SenseCam. Gemmell, Aris & Lueder show a small application to replay the SenseCam images of an interesting trip [59]. Due to its passive capture nature, this device may be liked by tourists as no effort is needed to capture images, thus allowing individuals to more fully enjoy their tourist trip. Hodges *et. al.* [71] detail the potential benefits of a personal visual diary such as that generated by the SenseCam. In preliminary experiments they have found that the use of a SenseCam dramatically aided a female subject, with limbic encephalitis[1], to recall events that happened during her day when reviewing SenseCam images. From our own personal experiences there is a noticeable improvement in the short-term memory recall of activities experienced during days while a SenseCam device was worn.

In fact the SenseCam has now become the prevalent lifelogging device with regard to memory focused studies. Investigations on the benefits of visual lifelogging to the human memory are now being carried out in the University of Exeter, the Medical Research Council in Cambridge, the University of Leeds, The Oliver Zangwill Centre for Neuropsychological Rehabilitation in Ely (all U.K.), Duke University, Columbia Medical School, University of Illinois Urbana-Champaign (all US), the University of Toronto, and the Hal-

[1] A form of encephalitis which is an acute inflammation of the brain

var Jonson Centre for Brain Injury Ponoka in Alberta (both Canada). All of this research activity is making use of the SenseCam device[2]. This means that in essence the contributions made in this thesis are of value not only to the lifelogging community, but also to the cognitive psychology community.

1.2 Motivation Scenario

In this section we will now detail 2 best-case motivation scenarios, envisaging how users will interact with their own captured visual lifelog data. The first case involves how an individual with a neurodegenerative disease can be aided to help remember a recent family event. The second case involves a tourist reviewing images of a memorable trip.

1.2.1 Memory Aid Scenario

Consider this scenario as a potential benefit that may be offered by lifelogging. Mary was healthy throughout her whole life, but unfortunately in her early sixties she has developed a memory impairment. This has been very distressing for Mary as she now very much struggles to remember events that happened as recently as the previous day. Until recently she has been embarrassed to talk to others for fear of forgetting seemingly trivial pieces of information, e.g. who she met at church 3 days ago, etc. However, recently she has begun to wear the SenseCam, a visual lifelogging device. She now puts it on at the start of every day after letting it charge overnight when sleeping. Every evening she uploads images to her PC, and our software automatically processes this data. As Mary's SenseCam takes approximately 2,000 images per day, it would overwhelm herself and her husband John to look through all these images. Therefore we automatically segment sequences of the images into distinct events/activities, e.g. when she had breakfast in the morning, when she went for a walk in the afternoon, when she was talking to Ann next door, etc. This helps herself and John more easily review what Mary has done in any given day. Indeed they particularly enjoy reviewing their Grandson's 4^{th} birthday that took place last Friday. Naturally there

[2] http://research.microsoft.com/sensecam/memory.htm

were some routine/mundane events in the morning first of all, so our software makes these events appear smaller, while the main event of the birthday gathering is magnified so as to catch John and Mary's attention. By playing through the images of this event Mary can remember that she was talking to her friend Jack about how quickly her Grandson is growing. For certain events John can prompt Mary on what she was doing, with the SenseCam images being very beneficial. Indeed Mary wonders what other times recently she was talking to Jack, so our system provides her with the ability to retrieve other similar events to the one of talking to Jack. She is then provided with a list of other potentially relevant events, which she quite enjoys looking at as they trigger some memories.

1.2.2 Tourist Scenario

Consider the following scenario as a motivation for the role of lifelogging. Aiden decides to go on a one day holiday to Dublin city centre and he wears his SenseCam all day long. That evening he wishes to review the SenseCam images of his day. Firstly it will be important for his viewing application to automatically identify the different activities/events he had for that day, e.g. in the morning he has breakfast at home, then gets the bus into the city centre, walks around O'Connell St., then gets a meal, etc. However, some of those events would be more important than others. Aiden has breakfast every morning, therefore its an unimportant event to him because it occurs regularly. However, while in the city centre he started talking to his friend John and this was a very important event due to its uniqueness as Aiden had not talked to John for quite a while, especially in Dublin city. That evening Aiden wonders when he was last talking to John, and he uses the application to find previous occasions similar to this. He can see that he had last talked to John 1 month ago, while it was almost 3 months ago since he talked to John in O'Connell St. However, he also nostalgically notices that he was talking to his friend Mary in O'Connell St. 2 weeks ago, as this event is visually quite similar to that of talking to John earlier today. While reviewing the event of talking to his friend Mary, he notices that he was at a football match later that evening. After browsing through some past memories Aiden then looks at his favourite picture from his trip today; an image of Trinity College. However, with the SenseCam images being of a low quality, quite understandable given that it is a passively capturing wearable camera, Aiden would like to get other pictures of Trinity College. Therefore the system then automatically retrieves images from numerous Internet sources that are taken in the same area as Trinity College, which provide additional memory retrieval cues for Aiden to reflect happily on his trip. Aiden is very amused to look at these high-resolution pictures, taken by other people, which are very similar to what he experienced earlier today.

1.3 Problem of Reviewing Lifelog Data

We have introduced what lifelogging is, and then focused on a visual lifelogging device, the SenseCam. We have also listed possible uses of such a device through ideal case motivation scenarios. However, we will now detail the difficulties in designing effective and automated lifelogging applications.

The majority of past research in the visual lifelogging domain has covered hardware miniaturisation [94] and also storage of images [60]. However, these challenges have been relatively well solved [71] in terms of better wearable devices, and in terms of inexpensive storage [15]. The challenge is now that of retrieving relevant information from the vast quantities of captured data. [15, 7, 89, 87].

Extensive research has taken place on the management of personal image collections in Dublin City University and elsewhere [116, 73]. However, these applications only consider pictures taken with current state of the art digital camera and mobile phone technology. With passively capturing wearable cameras, like the SenseCam device, the scale and frequency of images is much more significant. This brings us different challenges to those posed by current personal photo management applications.

One method to review images captured by the SenseCam is to use the SenseCam Image Viewer [71]. 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 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 different images representing encountered activities or events, coupled with the ability to search for events or similar events, could provide a much more useful method to manage SenseCam images. Lin and Hauptmann state that *'. . . continuous video need to be segmented into manageable units . . . '* [89]. A similar approach is required with respect to a lifelog collection of recorded personal images or video, and in Chapter 2 we will motivate in more detail why the approaches we have taken exploit characteristics of the human memory system.

Previous research in this area has been carried out on very limited datasets. No groups have captured data for the duration of an entire day over an extended period of time. Using the Deja View Camwear [123] Wang *et. al.* in their work state that *'. . . one of the authors carried the camwear, and recorded on average 1 hour of video every day from May to June . . . '* [155]. Similarly Lin and Hauptmann record data for only between 2 and 6 hours on weekdays [89], while others only capture for small time periods in the day too [146]. For all the experiments in this thesis we organised the collection of approximately 2 million lifelog images, passively captured by 12 wearers over a period of two years. 6 of those users have worn the SenseCam for at least one month, at an average of almost 10 hours per day. This means that we have collected a larger and more diverse dataset than other research groups, thus adding a stronger degree of significance to the results reported in this dissertation.

1.4 The Challenge

Even though we believe that visual lifelogging technologies, such as the SenseCam, will be very helpful as memory aids, we recognise that considerable difficulties exist in managing and retrieving lifelog data. We now specify four facets that we believe are important in terms of addressing this considerable challenge of managing lifelog data. These facets are: 1)

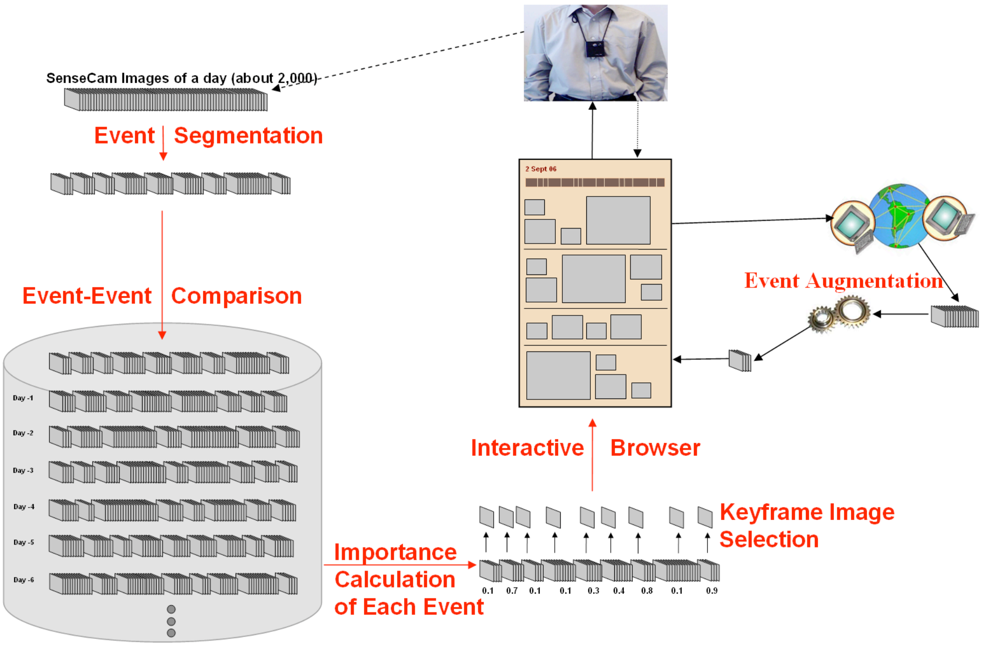


Figure 1.3: An overview of the challenges that this thesis addresses. Images are segmented into distinct events, which then enable those events to be compared against each other. By comparing events against each other we can then determine an importance score for each and every event. A keyframe must be selected for each event, which can also be augmented with material from external sources of information.

Event segmentation, 2) Event Importance, 3) Event Retrieval, and 4) Event Augmentation as illustrated in Figure 1.3. We now detail why we believe these are important challenges to address.

Techniques used in the traditional video retrieval domain are not necessarily suited towards managing lifelog data, as Aizawa notes that '*. . . the result of capturing human activities is a huge amount of multimedia data. New applications, different from those of traditional multimedia processing of TV and movies, will appear . . .*' [1]. Lin and Hauptmann state that a lifelog of '*. . . continuous video need to be segmented into manageable units . . .*' [89]. A similar approach is required with respect to a lifelog collection of recorded personal images or video as the SenseCam captures 2,000 images on an average day creating a sizable collection of images even within a short period of time, e.g. over 14,000 images per week which equates to approximately 700,000 images captured per year. Therefore no one individual could ever hope to retrieve images of encountered activities from their lives just by browsing. The challenge is to create an application to manage this large collection of images which raises the issue of how to organise this large image collection into "manageable" segments that can be easily retrieved by users. As will be detailed later in this dissertation, the human autobiographical memory system generally stores episodes of one's life. This is why we believe sequences of images must be segmented into distinct events or activities, and this process must be performed automatically (see "Event Segmentation" to top left of Figure 1.3).

Even when a lifelog of SenseCam images is segmented into events, there will still be approximately 20 distinct events in a user's day, which translates to over 7,000 events per year. The human mind stores autobiographical memories in a linked or associative manner, whereby we think of other semantically/conceptually related memories to a given memory [9]. In all aspects of the lifelogging domain it has been recognised that retrieval of relevant content is an important and significant challenge, e.g. by Bell & Gemmell [15]; Wang *et. al.* [155]; Ashbrook, Lyons & Clawson [7]; etc. We also believe that users should be given the ability to retrieve other similar events to any given event, e.g. "what other times was I at the park?", "show me other events of poster sessions", etc. This process, which is illustrated as "Event-Event Comparison" on the left hand side of Figure 1.3, should automatically go through the user's entire lifelog of events to find relevant items to a query event.

As mentioned in the previous paragraph, a user will experience approximately 20 events every day. The human mind more strongly encodes distinct autobiographical memories [120]. Indeed in studying the potential benefits of SenseCam images for those experiencing memory difficulties, Berry *et. al.* note that users are not interested in reviewing routine/mundane/"boring" events [17]. Indeed they point to other memory studies which indicate that even people with intact memory struggle to remember low salience and familiar events as they are generally uninteresting [40, 32, 90]. In browsing through content, Tancharoen & Aizawa note that *'... content-based features from audiovisual data are necessary to detect the significant scenes from our life ... '* [145]. As a result we believe that it would be useful to highlight or place more emphasis on events that are likely to be more important or interesting to the user as they browse through their lifelog of images. This should be processed automatically, and it also means that less important events can have less emphasis placed on them, or they may even be not displayed to the user at all as illustrated by the "Interactive Browser" in Figure 1.3.

In 1945 Vannevar Bush recognised that the human memory *'... operates by association. With one item in its grasp, it snaps instantly to the next that is suggested by the association of thought ... '* [28]. As will be detailed in Chapter 2, human memory generally operates by associating linked items together, we feel that it may be useful to augment an individual's SenseCam images of an event with images, and possibly videos, from other external sources of information e.g. if a wearer attends a big football match, it may enhance their enjoyment and also provide cues to other memories of the match if they were able to easily view additional photos and videos of the event; similarly if someone is at the Eiffel tower in Paris, their experience of reviewing their SenseCam event may be enhanced by photos taken by others in that location, where even a photo of a nearby park bench may be relevant as it evokes memories of the user sitting there. This presents a considerable challenge in intelligently mining the Internet to return relevant content, as illustrated by the "Event Augmentation" stage to the right hand side of Figure 1.3.

1.5 Hypothesis

Due to the fact that visual encoding is a strong facet of the human memory system [24], we believe that a wearable camera such as the SenseCam will be potentially beneficial to provide strong cues to retrieve autobiographical memories. The SenseCam has become a widely used device not only in the lifelogging domain, but also in the domain of memory studies. However, there exists a substantial challenge in managing the large collection of personal information collected by this device. To realise this, images will have to be segmented into distinct events e.g. having breakfast, being in work, at lunch, etc. The next challenge will be to automatically determine which of those events are most important or unusual e.g. going to a football match will be more memorable than working in front of a computer. Thereafter it will be necessary to have the ability to search for similar events to an example event e.g. *'find me other times when I was talking to my friend John'*. The final challenge will be that of augmenting the low-quality images from the wearable camera with higher quality images from external data sources, e.g. after attending a football match in Wembley the user will have the option to augment those images with location-tagged images from an external source like the Flickr database (containing 65 million geotagged images at the time of writing, but owing to its content being driven by a very large community of users, this is expanding by approximately 1 million images every fortnight).

Given that much research has indicated the effectiveness of cue-based recall, and that visual images in particular offer strong cues, we believe that visual lifelogging can offer an effective memory aid. However, given the overwhelming number of images captured by visual lifelogging devices like the SenseCam, the hypothesis of this dissertation is that:

> *In order for visual lifelogging technologies to effectively provide autobiographical memory retrieval cues it is necessary to automatically:*
>
> - *effectively segment a large collection of personal images into distinct events*
> - *accurately determine which of those events are most unusual*
> - *usefully provide the ability to link similar events together*
> - *augment images from the low-resolution wearable device with higher quality images from external data sources.*

Later in this dissertation we will anecdotally review the effectiveness of the *cues* provided by our automatic techniques. However, before that we will address each facet of the hypothesis individually. The structure of this thesis is organised as follows:

Chapter 2 gives an overview how technology may exploit certain aspects of the human memory system on occasion. A history of lifelogging research will then be presented, with a particular emphasis on visual lifelogging. Thereafter the technical intricacies of other researchers attempting to address the challenges of managing a visual lifelog will be explained, and also how the work carried out in this thesis differs to the prior work carried out by these other researchers.

Chapter 3 provides a detailed explanation of how sequences of images should be effectively segmented into distinct events or activities. The extensive experimentation and evaluation of this approach is also explained, detailing the areas of particular strength.

Chapter 4 addresses the challenge of accurately determining the similarity between any given lifelog events. Again extensive experimentation is carried out to evaluate the accuracy of the proposed approaches.

Chapter 5 proposes and evaluates the introduction of novelty as a useful technique to help determine the most important events in a given day of lifelog data. This chapter also explores the optimal method for selecting a keyframe image from a lifelog event.

Chapter 6 introduces the idea of augmenting SenseCam events with images/videos taken by many other people which have been made available on the Internet. This chapter proposes a number of methods to realise this idea, and extensive evaluation is carried out.

As the previous chapters have concentrated on technical evaluations of algorithms, Chapter 7 investigates the usability of a complete real lifelog browsing application which incorporates work from the previous four chapters. This chapter will in essence test the hypothesis of this dissertation by providing focused retrieval cues to end users. The effectiveness of our automated techniques in exploiting aspects of the human memory system to provide retrieval cues will be anecdotally reviewed.

Finally Chapter 8 concludes on the value of this work to the visual lifelogging domain, while also suggesting further areas of research to be undertaken in the future.

1.6 Contributions of Dissertation

The major contributions resulting from the work carried out towards this dissertation include:

- A review of cognitive science literature to investigate how best to exploit aspects of the human memory system in helping users review *their* own lifelog data.
- An improved event segmentation technique which only uses motion based sensors. Other important discoveries in using *TextTiling, peak scoring* and *thresholding* techniques are detailed in Chapter 3.
- Only 30% of image features need to be extracted for effective retrieval performance. Extensive evaluations are also carried out on normalisation and fusion techniques in Chapter 4.
- We discover that the introduction of novelty detection to the lifelogging domain is effective in identifying routine events, which are generally less interesting (Chapter 5).
- We have identified that an *image quality* keyframe selection technique is helpful in those events containing a large amount of visual change. In other events the simple approach of selecting the middle image in an event as the keyframe is sufficient (Chapter 5).
- We have introduced the idea of augmenting one's lifelog with images/videos from the Internet so as to provide additional memory retrieval cues. To achieve this we

introduce a technique to automatically construct textual queries by only having prior location information.

- Software distributions of work carried out towards the realisation of this thesis have been released to approximately 10 research institutions working with SenseCams.
- This dissertation has resulted in a number of peer-reviewed publications [50, 48, 47, 49, 46]

Chapter 2

BACKGROUND

Initially this chapter will provide a high-level overview of various investigations in cognitive psychology literature on how the human memory system functions. We will show that visual images are an excellent memory stimulant, and that in particular humans can retrieve past memories more easily when we are provided with visual cues towards those memories. This then motivates our desire to use visual lifelogging as being of potential benefit as memory aids. This chapter investigates work carried out by others in the domain of lifelogging, and gives a clear indication of the contribution of the work of this thesis to the domain. Then more general lifelogging work is described, followed by a particular emphasis on the visual lifelogging community. Thereafter some of the practical benefits associated with lifelogging are discussed. The important challenges that this domain needs to address are summarised, and finally detailed reviews are carried out on other event segmentation work as well as work on retrieving similar events, determining the most important events, and augmenting low-quality SenseCam images with high-quality external images.

2.1 Literature Review of the Human Memory System

We will now present a high-level overview on findings from the cognitive psychology field on the human memory system. It is our intention to motivate how various aspects of the human memory system can be exploited so as to aid us to more easily retrieve thoughts of past experiences from memory.

2.1.1 Layout of the Human Memory System

Atkinson and Shiffrin proposed a multi-store of memory in 1968 [8], as illustrated in Figure 2.1. The model states that there are 3 distinct memory systems present in the human brain, namely one to deal with sensory inputs, one that relates to short-term memories, and finally a distinct system that deals with long-term memories only. Since 1968 more advanced, and complex, memory systems have been proposed. In 1994 Schacter and Tulving summarised that there are five different memory systems [128]: **short-term memory** where we process a small number of items in working memory; **procedural memory** where we learn skills such as riding a bicycle; **semantic memory** which is knowing facts such as Paris is the capital of France; **episodic memory** which is the encoding of knowledge about personal or autobiographical memories (i.e. event memories); and the **perceptual representation system** which isn't just one system but several for the various perceptual modalities (e.g. visual representation system, auditory representation system, etc.)

We feel that the Atkinson-Shiffrin model provides a better starting point to grasp the basic functioning of the brain with regard to processing memories and we will now explore the three memory stores described in this system [8].

2.1.1.1 Sensory Memory

As we just explained Atkinson & Shiffrin proposed that there are 3 distinct memory systems within the human brain [8]. We now provide an overview on the first aspect within the Atkinson & Shiffrin model, namely the sensory memory system. In an influential book on the human memory system, Baddeley describes the sensory memory system is *"... intimately involved in our perception of the world ... "* [9]. In essence the sensory memory store only stores information for fractions of a second, and for each of the senses e.g. visual, audio, touch, smell, etc. Baddeley feels that this memory store, while important,

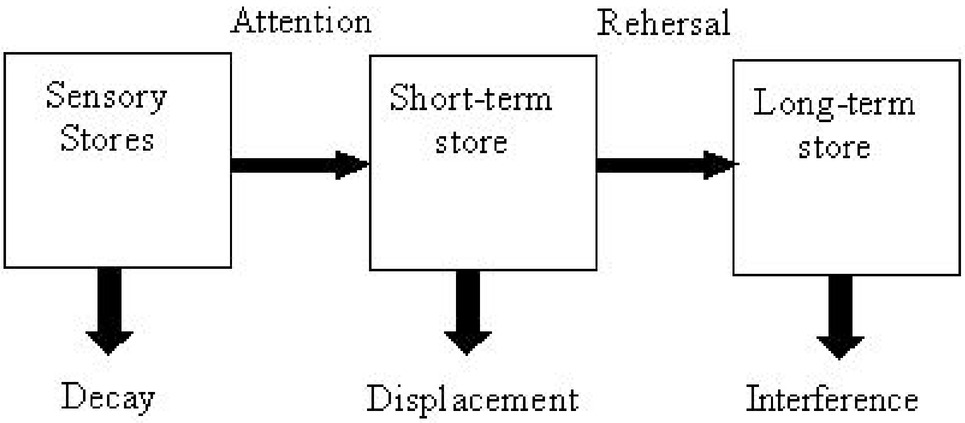


Figure 2.1: The human memory architecture as proposed by Atkinson & Shiffrin *(source: Elsweiler [54])*.

"... is probably best seen as part of the process of perception ... " and he also notes that we are able to measure a wide variety of senses of information (smell, touch, vision, temperature, etc.) but our brain only stores a small number of these senses [9]. This indicates that our brain highlights senses that are more important and chooses to store those.

2.1.1.2 Short-Term Memory

Having commented on the sensory memory system, we will now comment on the second memory system in the Atkinson & Shiffrin model, the short-term memory system. Elsweiler notes that much early work on short-term memory was concentrated on determining the amount of information that could be stored, how long it could be stored for, and why the information was then lost from the short-term memory system [54]. Indeed a seminal paper in the area is that of Miller in 1956 suggesting the capacity of short-term memory was 7 ± 2 items [101]. Although suggesting that short-term memory is of a fixed capacity, Miller explains how the process of "chunking" allows a greater amount of information to be stored e.g. the character sequence "b - i - k - e" is encoded as a single item/word "bike". Purdy *et. al.* point towards more recent research which has indicated that the capacity of short-term memory isn't limited to a fixed number of items [120], but rather to how much a person can pronounce in approximately 1.5 seconds as proposed by Schweickert & Boruff

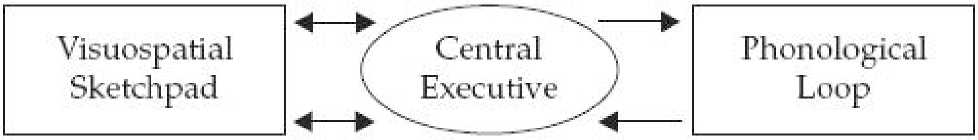


Figure 2.2: The Working Memory Model as proposed by Baddeley & Hitch *(source: Elsweiler [54]).*

[130]. Indeed Naveh-Benjamin and Ayres conducted an experiment to test how many digits English, Spanish, Hebrew, and Arabic speakers could remember; it was found that English speakers could remember just over 7 digits, whereas the Arabic speakers remembered less than 6 digits, owing to the fact that there are more multisyllabic digits in Arabic than in English [107]. Regardless of the precise amount of information that can be stored in short-term memory, Elsweiler points out that it is only retained for between 0.5 and 4 seconds, and he points to a number of theories ranging from *"... decay - forgetting due to decay of unused information ... and interference - forgetting because of new information interfering with old information ... "*, but this debate remains unresolved [54].

However, Baddeley states that short-term memory doesn't represent just one system, but a complex set of interacting sub-systems referred to as *working memory* [9]. Baddeley and Hitch argued that the term *short-term memory* should be replaced by *working memory*, which is a more thorough concept that considers an element of processing in addition to memory storage [10] as illustrated in Figure 2.2. Purdy *et. al.* make it clear that *"... Baddeley's model should be considered as an expansion of the Atkinson-Shiffrin model; it does not contradict or replace it ... "* The "working memory" framework proposed by Baddeley and Hitch consists of three components: 1) **Phonological Loop** (holds information in an acoustic format); 2) **Visuospatial Sketchpad** (holding information in a visual format); and 3) **Central Executive** (which is responsible for various control functions) [10].

Indeed Purdy *et. al.* have pointed towards past evidence suggesting humans translate visual stimuli into acoustic codes for short-term processing which supports evidence of the phonological loop existing [120]. However, they acknowledge this is not always the case as deaf individuals use visual codes during short-term memory processing, and also they point to studies carried out by Posner and his colleagues which suggest that participants sometimes rely on visual codes for representing short-term memories which provides evidence for the existence for the visuospatial sketchpad in memory [119].

In this particular subsection we have shown that short-term memory has a limited capacity, and that items decay quickly from this memory system. More recent researchers refer to this memory system as the *working* memory system. More information can be stored through the process of "chunking" information items, e.g. a sequence of letters chunks into a word, or a sequence of images could be chunked into a single image explaining the story of those images.

2.1.1.3 Long-Term Memory

Having described the sensory and short-term memory systems from the Atkinson & Shiffrin model, we now explore that last memory system proposed in their model, the long-term memory system. Atkinson and Shiffrin stated that the long-term memory system retains data for periods longer than 30 seconds and they speculate that it is unlimited in capacity [8]. Baddeley states that *"... of the three types of memory - sensory memory, working memory and long-term memory - the one that corresponds most closely to the lay person's view of memory is long-term memory. This represents information that is stored for considerable periods of time ... "* [9]. Anderson divides long-term memory into declarative and procedural memories [6]. Elsweiler notes that procedural memories refer to remembering skills e.g. how to ride a bicycle, how to walk, how to swim, etc. [54]. Anderson further splits up declarative memories into semantic and episodic memories, which are also identified by Schacter & Tulving [128]. We now explore each of these in a little more detail.

Semantic memories: Semantic memories are knowing about facts e.g. Paris is the capital of France, England won the soccer world cup in 1966, China is the most populous country in the world, etc.

Episodic memories: Episodic memories refer to those that are of personal experiences e.g. remembering when one's child first talks, recalling a conversation with one's friends from the previous evening, etc. Conway takes a slightly different position arguing that episodic memories are short term (in the order of just minutes or hours) e.g. remembering you closed the fridge door, that you made coffee, etc. He states that episodic memories are a bridge between working memory and autobiographical memory [39], *"... we conceive of episodic memory as a system that contains experience-near, highly event specific, sensory-perceptual details of recent experiences- experiences that lasted for comparatively short periods of time. These sensory-perceptual episodic memories do not endure in memory unless they become linked to more permanent autobiographical memory knowledge structures, where they induce recollective experience in autobiographical remembering ... "* Meanwhile Tulving describes that *"... episodic memory does exactly what the other forms of memory do not and cannot do - it enables the individual to mentally travel back into her personal past ... "* [149].

Purdy *et. al.* believe that the case of a famous test subject, H.M., is a strong indication that short-term memory and long-term memory are quite distinct from each other [120]. They detail the critical role played by the hippocampus[1] in memory using the case of a subject (H.M.) who had severe epileptic seizures, and surgeons attempted removing the hippcampi from his brain. The surgery relieved the epileptic seizures, but H.M. then had an inability to learn new events. *"... he had some trouble remembering information from one to three years before the operation, but little trouble remembering older events ... he would read the same magazine repeatedly without any loss of interest ... he had enormous difficulty learning new facts and remembering new events, but he has little difficulty learning new skills ... "* Maguire, in a neuroimaging analysis, also highlights the role played by the hippocampus which is activated in retrieving episodic memories [93]. In some instances an impaired hippocampus may leads to severe memory impairments such as amnesia. In fact Baddeley also notes that Alzheimer's disease accounts for over 50% of cases of de-

[1] The name of a particular region in the brain. There are 2 hippocampi in the brain, both are approximately symmetrical to each other in terms of their location on the brain.

mentia, and occurs in 10% of people over 65. The likelihood of having Alzheimer's disease increases with age [9].

To summarise this particular subsection we have detailed aspects related to the long-term memory system in the Atkinson & Shiffrin model. According to Schacter & Tulving this memory system can be broken up further into procedural, semantic, and episodic memories [128]. Episodic memory is of particular interest to us, as it is responsible for how we store our autobiographical memories e.g. remembering what you did on your 24^{th} birthday. We have finally pointed towards evidence stating that there are many people in the world who have difficulties in accessing their long-term autobiographical memories.

2.1.2 Importance of Cues to Aid Retrieval of Items in Memory

There is a body of research indicating that providing people with memory cues, that are quite similar to the way the particular memory is encoded in the brain, will lead to enhanced retrieval of memories. Purdy *et. al.* state that *"... scores on recognition tests are usually much higher than scores based on recall ... "* [120]. This indicates that users prefer to be prompted with cues, rather than being asked to retrieve memories from scratch. Baddeley also echoes this point, stating that *"... it is as if the cues direct you to search in the appropriate location in memory, and as such allow access to traces which would otherwise have been missed. While there is no doubt that both cueing and recognition can reveal information which is not accessible using straightforward unaided recall, it could also be argued that the memory trace is present but not strong enough to allow recall ... "* [9]. In essence good cues, which are related to how the brain encoded the memory in question, lead to strong retrieval performance.

Elsweiler comments in great detail on the cue-dependent forgetting theory introduced by Tulving in 1974, *"... which argued that information is available in memory but cannot be accessed without the appropriate "cue" ... "* [54]. Tulving later developed this theory into the encoding specificity principle, in which memory performance is dependent on the similarity between the information in memory and the information/cue available at retrieval time [148]. Elsweiler also points to other studies with results endorsing Tulving's theory on strong cues leading to better recall [54]. Indeed in a study of memory about office activities in a desktop computing environment, Czerwinski & Horvitz also found that cue based recall is much more effective for retrieving memories than free recall [44].

Finally Elsweiler details that the discriminating value of a cue impacts its effectiveness in retrieving memories in the brain, *"... people have been shown to falsely remember words based on their similarity e.g. if given a list of cake ingredients to remember such as flour, sugar, butter etc., people will often claim to have remembered eggs appearing in the list. This is known as the Deese/Roediger-McDermott effect [126]. This effect can be minimised by providing discriminative cues - cues that ensure that participants utilise specific recollections rather than relying on general familiarity ... "* [54]. However, Schacter *et. al.* have shown that images are effective discriminative cues, through presenting subjects with images along with words in the list of cake ingredients, leading to better discrimination for the ingredients present [127].

2.1.3 Importance of Encoding towards Memory Retrieval

We have just motivated that cued recall is more effective than free recall in aiding people to access autobiographical memories, however it is now worthwhile to consider what the most effective cues that can be provided are. It is interesting to note research indicates that it is easier to retrieve experiences from memory when the retrieval cue is quite close to the encoding cue. Godden & Baddeley carried out a study on teaching new words to scuba divers on land and under water [63], and discovered in the words of Purdy *et. al.* that *"... performance was relatively poor, however, when participants learned and recalled under different conditions. Thus, making the conditions at retrieval (underwater) the same as those at encoding (also underwater) led to good retrieval ... "* [120]. Purdy *et. al.* indicate that memories initially decay very quickly, but then more slowly over time [120], in reporting a study on the retention of mathematical skills learned by a class, they *"... still found measurable levels of retention 50 years later. In fact most of the forgetting appears* *to take place in the first few years after the material was learned . . . retention in long-term memory, even for intervals up to 55 years, are a function of initial encoding . . . "* These studies indicate that proper encoding is important to aiding retrieval of memories from the brain.

Visual encoding is strong: Indeed Brewer has found that more than 80% of randomly sampled memories consisted of visual images [24]. To further reinforce the notion that images are important to memory retrieval, Shepard has found that people have a much better memory of pictures presented to them than merely a list of words [134]. It is our belief that this recognition of images will be even stronger if the images presented to people are actually images of personal meaning to them, as it has been reported that the "generation effect" indicates that people remember things better if they generated those things themselves [66, 54]. Indeed Conway believes that autobiographical memories are very often in the form of visual mental images [39].

Remembering vs. knowing: Gardiner describes 2 distinct elements of autobiographical memory, autonoetic (self-knowing) consciousness and noetic (knowing) consciousness. Autonoetic consciousness *". . . is the mental reinstatement of personal experiences of previous events at which one was present. Noetic consciousness is expressed without any such self-recollection but simply in awareness of familiarity, of knowing . . . "*. Berry *et. al.* [17] have indicated that visual images taken of one's life may be helpful in terms of "remembering" an episode in their life, commentating that, *". . . retrieval of a visual image distinguishes autobiographical memory from autobiographical knowledge, that is, visual images enable a person to recollect or relive an event, rather than simply know of an event . . . "*.

Encoding viewpoint is important: Burgess *et. al.* speculate that people with episodic memory impairments may be hampered in recognising happenings from a different viewpoint [27]. They report on a study carried out by Vargha-Khadem *et. al.* on a patient, Jon, with a focal bilateral hippocampal pathology [151]. In this study they used a virtual reality system to present objects located in three-dimensional space and tested recognition memory for their locations. It was noted by Burgess *et. al.* that *"... Jon's enormous impairment when the viewpoint was changed but only mild impairment from the same viewpoint strongly indicates a role for the hippocampus in storing allocentric representations of object locations ... "* We believe that it may be helpful for people with such impairments to have visual images taken from their perspective/viewpoint as retrieval cues, as they would be encoded in a similar fashion to those autobiographical memories recorded in the brain.

Memories can be temporally encoded: Larsen, Thompson, & Hansen note that *"... people are quite accurate in judging the time of events, that is, temporal judgements are normally unbiased estimates of actual time in the past ... "* [86]. Essentially people are quite good at narrowing down when particular autobiographical memories happened to a specific time period, which could be a day, week, month or year. Davies & Thompson believe that people use extraordinary events as anchors when trying to retrieve memories from the past, although people still do have a good sense of the temporal ordering of those memories [45].

Encoded memories are linked by association: Baddeley comments on the richness and flexibility of how memories are stored in the human brain, stating that *"... memory is concerned with concepts or ideas; which are in some cases clearly related to words but are not in themselves words ... "* [9]. The meaning of this is that when we think of a certain autobiographical memory, our mind finds associations with other long-term memories that are conceptually similar to the given autobiographical memory in question. Indeed Elsweiler notes that the storage of *"... our memories of experiences, events and stories are determined not only by the story itself, but by our background knowledge ... "* [54] meaning that we try and associate new experiences with semantic and episodic information already stored in memory.

Distinctiveness is important in memory encoding: Purdy *et. al.* argue that the notion of distinctiveness is very important in memory encoding [120], stating that *"... the more distinctly a memory is coded, the easier it will be to recall that memory later ... "*. They report the Von Restorff effect which shows that even nonsense syllables become more memorable than meaningful words when the nonsense syllable is the isolated item on the list. This notion of distinctiveness is very important, and we believe that those memories that are most interesting for users to review from their own lives, are the events which were most distinct to the user e.g. talking to a new person, going to a new place, etc.

To summarise this subsection we have detailed the way in which memories are encoded in the brain impacts on the retrieval of those memories/experiences at a later time. We have shown that visual images are important to memory retrieval, and in particular help people "remember" an event better than simply "know" about it. Also we have looked into the fact that for people with severe memory impairments, they may only encode autobiographical memories from just one fixed viewpoint. We have pointed towards research showing that in general people are quite accurate in judging the time of events, while also indicating that memories are stored in an associative fashion in the brain. Finally we have made it clear that the notion of distinctiveness is very important in memory encoding.

2.1.4 Considering Technology to Aid Memory Retrieval

Thus far we have given a high-level overview of how the human memory system functions. There are 3 major categories of memory system: sensory, short-term, and long-term. We have shown that short-term memory has a limited information capacity, although this can be somewhat increased through the process of chunking. We have shown that there are two type of long-term memory, semantic memory (e.g. knowing London is the capital of England) and more interestingly episodic/autobiographical memories (e.g. remembering events that you personally experienced). Conway has argued that these autobiographical memories are very important and in fact define who we are and how we act [40]. We have shown that cued recall leads to better retrieval of memories from the brain than free recall, i.e. good memory cues are important. Better retrieval cues are those that most closely resemble the way that the memories in question were initially encoded in the brain. We have also detailed how such memories are mainly visually encoded in the brain. The visual viewpoint in which memories are encoded is also an important feature on how episodic memories are encoded, as is how distinctive the memories are.

However as we pointed out there are a growing number of people with memory impairments, e.g. 10% of people over the age of 65 have Alzheimer's disease [9]. Indeed Schacter notes that while biological memory generally serves us well, it is highly selective and fallible [126]. Indeed many times people comment that they wish they had better memory, and this is perhaps what has driven people towards maintaining written diaries, i.e. to help people remember what they did. Naturally given that diaries take much effort to maintain, the idea of exploiting technology to aid memory has come into focus. Barreau states that *'...Digital memories surpass biological memories by enabling people to directly see and interact with captured information. This powerful capability can help people manage their personal information ... '* [13].

In fact some recent research making use of lifelogging images has shown much promise in providing good cues to retrieve autobiographical memories [71, 17, 131, 87]. Indeed we shall now provide an overview of the origins of the lifelogging field, then some reasons as to how it can be a possible solution in providing effective memory cues, and finally the challenges that must be solved so as to make this technology usable.

2.2 A History of Lifelogging

Pictures, books, monuments, and storytelling were all used so as to leave a "legacy" of oneself or of a group of people. Human memory is far from perfect and throughout history we have devised many techniques to help us remember what we did. Traditionally the written diary has been a very popular means for people to record events that have happened in their lives. For much of this time, diaries have been used almost as a form of release for individuals to get inner most thoughts "off their chest", however still a large number were used as a form of record for important happenings in a person's life, e.g. weddings, births of children, a son/daughter leaving home, etc.

In 1917 Buckminster Fuller, an American scientist, was *"...determined to employ my* *already rich case history, as objectively as possible ...* ”[2], and so he began recording every minute detail of his life in diary format. For the rest of his life, and 500 volumes later, he had an extensive record of all his lawyer bills, medical documents, plus written descriptions of much of the minute details of his life. However, this required a significant effort to not only capture information, but also to retrieve it. Vannevar Bush, who coordinated the activities of approximately 6,000 scientists in the application of science to warfare, recognised the potential benefits of automatically managing the overwhelming amount of information and events that we encounter [28]. In a visionary paper in 1945 outlining where the future focuses of scientific researchers should lie after the 2nd World War effort, he envisioned a “Memex” device which would extend the human memory, and which would be a *‘... device in which an individual stores all his books, records, and communications, and which is mechanized so that it may be consulted with exceeding speed and flexibility ... ’.*

For quite a while the idea of managing personal information wasn’t entertained until the advent of personal computing, and in particular until the ability to store a very large amount of data was possible. Steve Mann, now a researcher at the University of Toronto, spent much time, from the 1970’s onwards, trying to capture much of what he saw through the design of head-mounted video cameras as illustrated in Figure 2.3 [94]. His vision opened up a whole new, very rich, source of information which is of much benefit and addition to an individual’s lifelog.

However apart from these pioneers the field of lifelogging is a relatively new but rapidly expanding area of research. As noted by Gordon Bell and Jim Gemmell, *‘... in recent years, however, rapid advances in storage, sensor and processor technologies have paved the way for new digital recording and retrieval systems that may go far beyond Bush’s vision ... ’* [15]. They have noted that in 20 years time a 250 terabyte hard drive (*“capable of holding tens of thousands of hours of video and tens of millions of photographs”*) may well only cost $600, which should be enough to store all of the personal information encountered in an individual’s lifetime. Indeed O’Hara gives a good overview on what motivates us to

[2] www.bfi.org/node/105

investigate lifelogging activities stating that *"... Every piece of information is such that it is very unlikely, but just possible, that it is valuable. Before technology allowed comprehensive storage, our strategy was usually to try to estimate which information is likely to be more valuable and to keep that. Now there is no reason to stick to that philosophy ... "* [113] i.e. we don't want to delete anything!

Of these recent projects, perhaps the most extensive lifelogging effort is the MyLifeBits project which is attempting to relate many different sources of information together into one integrated framework to enable users search for multiple types of information at any given instance in time [61, 60, 15, 35]. The sources of information that they attempt to store and retrieve include: instant messaging text, web pages visited, e-mails received, telephone conversations, television programmes watched, radio listened to, audio recorded conversations, GPS tracklog, and also SenseCam images. All this data is stored in a large database, the MyLifeBits database. The various strands of information are associated through time stamps. They attempt to make real the vision of Vannevar Bush that many items should be linked together, because the human memory works by associating items together. As also postulated by Kelly & Jones the end-goal lifelogging system should incorporate many sources of information, such as: web pages, e-mail, GPS, Bluetooth, images, etc. [78].

The MyLifeBits project is very broad, and within this there exist many challenges, one such example being the management of audio recordings. Ellis & Lee have investigated the segmentation of audio data into distinct activities (e.g. in the house, on the subway, at the park, at work, in a meeting, etc.), and also the clustering and retrieval of audio events (e.g. find me other times I was in the park, other times I was at a play, etc.) [53, 52]. Vemuri has also investigated the retrieval of audio activities and recognises that recording everything is relatively straightforward, but searching for specific audio snippets is a considerable challenge [153]. Staying on the theme of audio-focused lifelogging efforts, Kern, Schiele and Schmidt address the challenge of classifying a person's activities based not only on audio activity levels but also on movement/accelerometer sensors, which were found to be quite useful [109].

Bristow *et. al.* in Birmingham University [25] have conducted a user study to define the different types of contextual information that are important in terms of wearable computing. A few of the highlighted types of context that were mentioned include: location, people, time, indoors/outdoors, physiological indicators, weather, and how frequently a task is done. The SenseCam captures time, movement data, possibly the presence of people (through passive infrared sensor), the temperature, and through our novelty analysis (described later) an indication of how frequently an activity has taken place; thus incorporating many of the contextual sources of information suggested by Bristow *et. al.* [25].

2.2.1 Focus on Miniaturisation and Storage

Much research in the past has concentrated on miniaturising visual lifelogging capture devices so as to encourage more users to become comfortable with this concept. Several research groups have had visual lifelogging devices that required users to wear a laptop carried on a bag around their backs [144, 89] and in some cases a head mounted camera [147]. As McAtamney and Parker note in their study, both the wearer and the subject talking to them are aware of personal recording devices while holding conversations [99]. Therefore it is desirable to decrease how noticeable a wearable device is so as to encourage more natural interactions with the wearer.

Steve Mann has pioneered the effort to make these visual lifelogging devices smaller [94, 95]. Figure 2.3 illustrates the work carried out by Mann over 3 decades to make visual lifelogging devices less intrusive. Blum, Pentland, & Troster focus on *when* an image should be captured/triggered in their “InSense” project [22]. They identify activities using trained classifiers on captured sensor data, and using this information they can identify a given activity in real-time. Therefore more pictures are taken during interesting activities, and images are captured less frequently during routine/mundane activities. The “InSense” system is a bit cumbersome in terms of its dimensions though (see pack in Figure 2.4).

Research is still ongoing in trying to make lifelogging devices even smaller and more ubiquitous. Given the prevalence of mobile/cell phones, the WayMarkr project of New

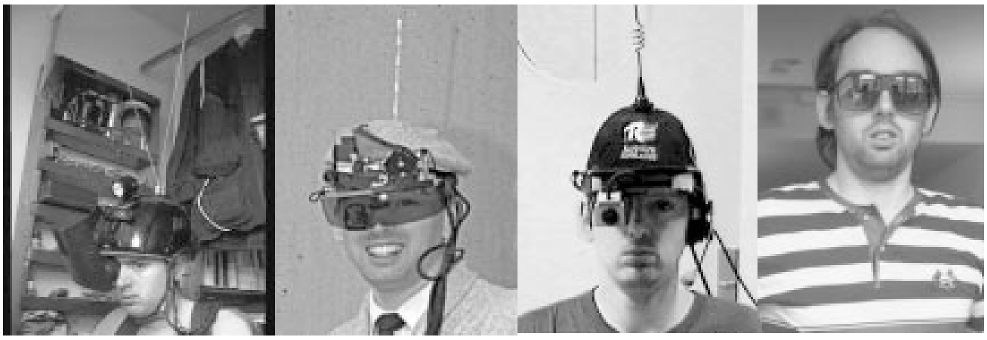


Figure 2.3: The evolution of Steve Mann's personal video capture devices since the 1970's *Source: Mann [94].*

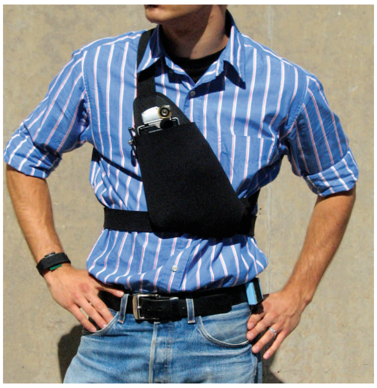


Figure 2.4: The "InSense" system developed at MIT stores data on a PDA contained within the pack *Source: Blum, Pentland, & Troster [22].*

York University uses a mobile phone affixed to a strap so as to take pictures automatically [26]. As it is a mobile phone they can extract cell tower ID for location information. The best feature of the project is that photos are uploaded to a remote data warehouse, therefore storage on the device is not an issue. In this project much focus has been on the capture and storage of information, as opposed to the retrieval and organisation of all the images. The DietSense project in UCLA also makes use of a mobile/cell phone, hung via a lanyard around the neck in a SenseCam like fashion, to capture pictures automatically [122]. The goal of this project is to assist medical doctors reviewing the dietary intake of patients so as to make more informed decisions in recommending personalised diets. However, Reddy *et. al.* do identify a significant challenge in automatically managing this form of captured lifelog data [122], and in fact we have worked in collaboration with the DietSense researchers at UCLA providing them with analysis tools to manage their lifelog collection, using techniques that will be mentioned in this thesis in due course.

Microsoft Research in Cambridge, U.K., has further advanced the field through the introduction of the SenseCam [71]. 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. The SenseCam holds advantages over video recorders as the device only takes images on average 3 times per minute, thus allowing a person to quickly review all the images to gist what has happened in a given day, rather than the requirement of watching a video clip in real time. An even bigger advantage is the fact that storage requirements are reduced, and also privacy concerns are not as grave as the camera takes snapshots as opposed to continuous footage. The SenseCam is now used by not only lifelogging research groups, but also by research groups investigating the benefits offered by lifelogging to those with memory impairments, as will be detailed in section 2.2.4.

Ashbrook, working in Georgia Tech, envisages that the lifelogging device of the future will be quite like the SenseCam but also with video capture and an LCD screen for reviewing images [7]. This device would also automatically upload images to a user's blog. However, this is a position paper, with no such implementation currently existing.

Given the large advances taken by the hardware industry in producing hard drives, the problem of storage limitation has disappeared. As stated earlier Bell & Gemmell predict that it will be possible for a person to store their whole lifelog of data for $600 worth of storage within the next 20 years. While there is still effort in terms of hardware miniaturisation, and increasing the lifespan of batteries, the lifelogging field are now on top of the challenge of recording/capturing the visual information that we encounter.

2.2.2 Shifting Towards Retrieval

Given that the majority of past efforts have been on the physical hardware of the lifelogging devices, it is only recently that as a community we have moved towards the very considerable challenge of making the content understandable and searchable.

Gemmell & Bell [61, 60, 15] and Tancharoen & Aizawa [3, 145, 146, 144, 72, 2] have produced many papers, and the common theme in all is that the biggest challenge facing the lifelogging community is that of efficient retrieval of information that is useful to the user. The work carried out in this dissertation is focused on the efficient management and effective retrieval of lifelog data that suits the information need of the user.

Extensive research has taken place on the management of personal image collections in Dublin City University and elsewhere [116, 73]. However, these applications only consider pictures taken with current state of the art digital camera and mobile phone technology. With passively capturing wearable cameras, like the SenseCam device, the scale and frequency of images is much more significant. This entails a different set of challenges to those posed by current personal photo management applications.

2.2.3 Privacy Concerns

As lifelogging has become more prevalent, so has the awareness of the potential benefits that it offers. However, some people have expressed some concern over the fact that many of the mundane details of our lives will be recorded, and they feel that there are some drawbacks to this.

For example Bannon in the University of Limerick points out that it's very important to forget things as well as remember them, *"... Forgetting is a central feature of our lives, yet it is a topic which has had relatively little serious investigation in the human and social sciences ... Forgetting is not some unfortunate limitation of the human, but is rather a necessary mental activity that helps us to filter the incoming sensory flood, and thus allows us to act in the world ... "* [12]. Nack in the University of Amsterdam has also raised concerns about lifelogging devices and their potential intrusiveness [104]. Both authors feel that these concerns outweigh the benefits offered by lifelogging devices.

Indeed, from personal experience of capturing an audio lifelog dataset, we experienced great resentment and unease by others [50]. Some work colleagues were initially uneasy about visual information being captured, but after time they came to accept or trust this source of information, even when in the hands of others. However, almost everyone was uncomfortable with audio information being recorded, and a great deal of suspicion and fear surrounded how such information would be stored. Bell & Gemmell in espousing their vision of lifelogging are very conscious in stressing that the stored information must be secure [15].

It is interesting to consider the views of someone outside the field of lifelogging, and Allen, a law professor, points out that lifelogging could fuel excessive self-absorption, *" ... since users would be engaged in making multimedia presentations about themselves all the time ... "* [4]. She also points out that lifelogging *" ... would extend the longevity of personal misfortune and error ... dredging up the past can hurt feelings, stir negative emotions, and ruin lives ... "*. She is also concerned about the secure storage of lifelog data. She concludes that *"... a counter-technology to block lifelog surveillance should be designed and marketed along with lifeloggers. The owner/subject of a lifelog should be able to delete or add content at will. No one should copy a lifelog or transfer a lifelog to a third party without the consent of its owner ... "*.

On the topic of lifelogging and privacy concerns, O'Hara makes an interesting counter-argument to the views of Allen, in expressing that *"... reconciliation has been possible in*

Germany and South Africa partly because of commemoration of the horrors of the holocaust and apartheid . . . " [113]. In fact Emma Berry of Microsoft Research and Addenbrooke's memory clinic in Cambridge, U.K., made an excellent point on a BBC Radio 3 memory program arguing that the SenseCam can aid forgetting at times, e.g. if ones grandmother had a severe case of Alzheimer's disease and had become quite violent in her last few years, one would perhaps struggle to remember the happy occasions with their mother . . . however if she had a SenseCam her whole life, it would be possible to look at those pictures of the times when she was healthy, this would influence one's thoughts on their grandmother better, and help them *forget* how violent she has been recently[3].

It is beyond the scope of this thesis to deal with the sociological aspects of lifelogging, however it is important that the reader is aware of such concerns being expressed by other members of the community. We are not of the opinion that everyone *should* be wearing lifelogging devices, however we feel that those who may benefit from wearing, and indeed rely on, such devices should be allowed do so.

2.2.4 Memory Benefits

While there are "invasion of privacy" concerns associated with lifelogging, and while these concerns must be addressed before this area becomes mainstream, it is important to acknowledge that there are many applications which will be of great benefit to a large number of users.

As Schacter notes, biological memory serves us well, but it is highly selective and fallible [126]. The goal of automated lifelogging capture and retrieval has come into focus. Barreau states that *'. . . Digital memories surpass biological memories by enabling people to directly see and interact with captured information. This powerful capability can help people manage their personal information . . . '* [13].

Microsoft Research and the Memory Clinic in Addenbrooke's Hospital in Cambridge, U.K., point out that lifelogging systems do not capture human "experience", rather they

[3] http://www.bbc.co.uk/programmes/b00dkvqw

capture a set of cues to help trigger the "remembering" of human experiences [131]. In one piece of work they explore 4 basic questions: 1. Do SenseCam images improve our memory? 2. Can people recognise images as their own? 3. How does the effect of remembering change over time? And 4. Are manually captured images better retrieval cues than passively captured ones (taking into account the generation effect mentioned by Slamecka and Graf [135] whereby we regard our own generated content as more memorable)? It is interesting to note that '*... in 95% of cases, when subjects reported any part of an event, they were able to report all the constituent parts ...* '. Regarding their findings as regards to the aspects explored in their work: 1. Does the SenseCam significantly help improve people's memories? YES; 2. Do people recognise their own images? YES; 3. How does the effect of remembering change over time? Over time we can remember more with the SenseCam than we normally would, however the power of the cues does deteriorate in terms of "remembering"[4], but in terms of "knowing"[5] it seems less sensitive to the passage of time - interestingly even before viewing images people had better memory of events when they wore the SenseCam than not, which may be due to them taking manual pictures and remembering doing that, or comments passed on it; And 4. Are manually captured images better retrieval cues than passively captured ones? Contradicting the generation effect, passive capture seemed to do better (not significantly so it was stressed) than manual capture in helping the subjects remember. In conclusion the results of this work indicate that lifelogging supports both the ability to remember and to know what happened in one's personal past, with the "knowing" being most effective.

Berry *et. al.* carried out a study to investigate the benefits of lifelogging to a 63 year old lady with limbic encephalitis [17]. In this study the woman had access to a SenseCam over a 12 month time period. She was able to review recorded events with her husband, who could help her retrace what they did together to assist her memory. The results of this small scale study are quite impressive as illustrated in Figure 2.5. Here it can be seen that the potential benefits offered by a visual lifelog can be better than the traditional written diary.

[4] can re-experience event in the "mind's eye"

[5] can't specifically remember, but other knowledge makes them "know" it happened

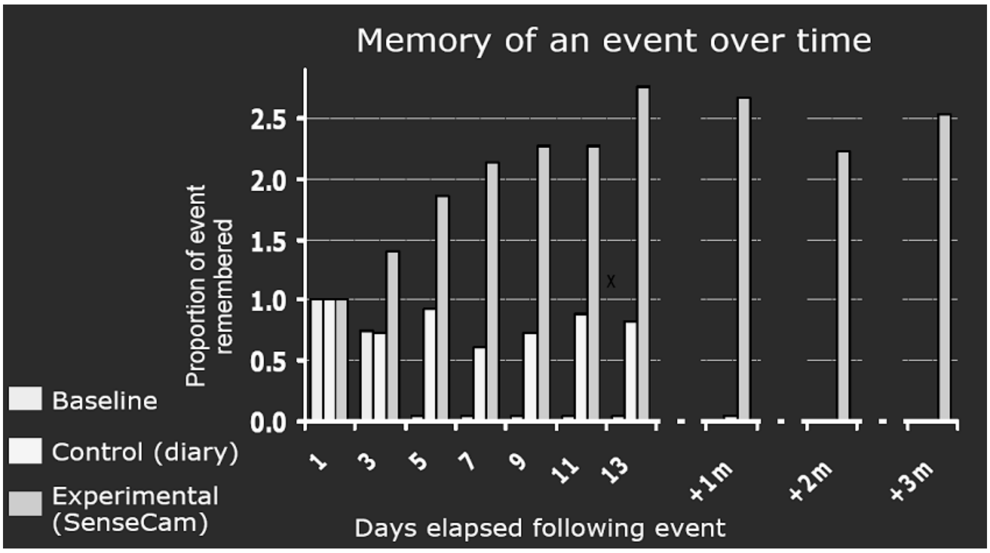


Figure 2.5: This graph illustrates the potential benefits offered by the SenseCam to those with memory impairments *Source: Hodges et. al. [71].*

In personal communication with Emma Berry and Georgina Brown of Microsoft Research and the Addenbrooke's Memory Clinic in Cambridge, UK. we have been informed that 4 others with memory impairments were given lifelogging devices, and the results again are promising in terms of their memory recall.

Harper *et. al.* also show that the SenseCam is a good aid to remembering past activities, in that it provides new perspectives on the past and users like to review the images too, finding them quite amusing to watch [69, 68]. They note that " *...what we find with devices like SenseCam is that, although they might not act as memory prostheses, they most emphatically don't miss the mark in offering memory-related experiences that appeal to users ...* " Harper *et. al.* also comment that people wearing SenseCams capture more "natural" photos, with one participant suggesting that " *...there is simply no other way of getting results of this kind ...* "

To further illustrate the suitability of the SenseCam towards work of this nature, Baecker *et. al.* are using the SenseCam to investigate the benefit of multimedia biographies to those with Alzheimer's [11]. Fleck is investigating the benefits of the SenseCam to school teachers [55]. Furthermore Greenberg *et. al.* are exploring the usefulness of the SenseCam as a technology to support students' memory and reflection [65].

Not all the benefits offered by lifelogging are just to aid the human memory. Lifelogging also offers other potential applications and benefits and a selection of these will now be discussed.

Ashbrook in Georgia Tech feels that automatically converting a lifelog into an online blog is a potentially significant application [7]. Given the large explosion of web blogs and also personalised "Web 2.0" social networking sites, this ability to automatically summarise one's day and make it available for friends online may well be very popular. He does recognise there are still big problems in efficiently browsing and searching the recorded lifelog data, which is what we are attempting to address in this dissertation.

Fleck & Fitzpatrick in Sussex carried out a study indicating that the SenseCam could be useful in terms of supporting reflective learning [55]. A study was conducted on students reviewing their performance at arcade games, and how SenseCam images help promote discussion and retain memories of the event.

Hodges *et. al.* note that *'...the act of stopping to take a photograph is very disruptive; it can break the patient's train of thought and it may be inappropriate from a social perspective'* [71]. Taking this into account, we have described a motivation scenario in the introductory chapter of this thesis whereby lifelogging devices can be useful tourist aids. The wearer of the device can relax and enjoy their holiday, without the burden and hassle of having to manually capture photographs.

2.2.5 Challenges in Aiding Memory Retrieval via Visual Lifelogging

As we noted at the start of this chapter, it is believed that autobiographical memories are mainly visually encoded in the brain, and the visual viewpoint that memories are encoded from can provide good cues to aid retrieval. The SenseCam records pictures from the viewpoint of the user, thus being able to provide recall cues that are very close to how the original memories/experiences were encoded in the user's brain. To quickly digest large volumes of SenseCam information, we believe that it will be useful to chunk sequences of

SenseCam images into distinct events or activities, thus allowing a user to quickly browse through his or her images. We detailed evidence supporting the fact that more unusual episodes are encoded strongly in the memory system, thus it will be desirable to draw the attention of the users to these events when reviewing their SenseCam images. Also given that we have described how the human mind organises memories in an associated manner, it will be useful to enable users to search for events/memories related to any specific event/memory that the user encounters.

A significant challenge is in managing the large collection of personal images that are generated by the SenseCam. This problem is made very difficult due to the age-old challenge of addressing the "semantic gap". Much effort has been made in the video and image research communities in addressing this difficult challenge [143]. As was stated towards the end of Chapter 1, the hypothesis of this dissertation is that:

> *In order for visual lifelogging technologies to effectively provide autobiographical memory retrieval cues it is necessary to automatically:*
>
> - *effectively segment a large collection of personal images into distinct events*
> - *accurately determine which of those events are most unusual*
> - *usefully provide the ability to link similar events together*
> - *augment images from the low-resolution wearable device with higher quality images from external data sources.*

Previous approaches used by others to address these challenges will now be discussed in the remaining subsections.

2.3 Event Segmentation Approaches

This section will review techniques used by others to segment sequences of lifelog images into distinct events or activities as illustrated in Figure 2.6 e.g. breakfast, working on PC, at a meeting, etc. Zacks, who studies how representation in the brain works, notes that *"... segmenting ongoing activity into events is important for later memory of those activities ..."* [158]. Staying in the field of cognitive science, Newtson, Engquist, & Bois note that *"... breakpoints* [between events] *tend to correspond to points at which the most physical features of the action are changing ..."* [108]. Indeed, in the lifelogging community it has been recognised by Lin & Hauptmann that a visual lifelog should be segmented into shots/activities/events to make it manageable [89].

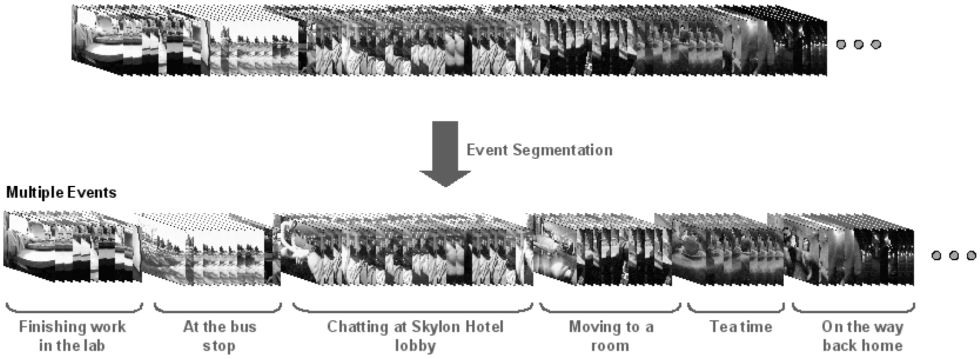


Figure 2.6: Segmenting sequences of lifelog images into distinct events/activities.

In the traditional video retrieval domain it has long been accepted that video should be organised into shots [136, 42]. Thus it is natural to segment a lifelog into distinct activities like the "table of contents" approach discussed by Xiong *et. al.* [156]. Video consists of a series of images usually taken at 25 frames per second; the relative difference between adjacent images from the same shot is thus quite minor unless there is huge movement of objects within the frame, or of the actual camera. Existing image analysis techniques for shot boundary detection use this information to compare visual similarity between adjacent frames and if similar, then assume the images are from the same shot. Exceptions to this also occur where there is a gradual transition in a shot bound, like a fade from one shot to another, but the above is a generic description of how shot boundary detection works. In video retrieval the performance of modern shot boundary detection systems is over 95% in terms of precision and recall, with respect to hard cuts [141]. However, when it comes to gradual transitions the performance is just over 80% which may indicate an upper bound for event segmentation performance with respect to the lifelogging domain. In the case of segmenting adjacent SenseCam images into events, these images are taken at a rate of up to 3 images per minute and the difference between adjacent SenseCam images is thus likely to be much greater than for adjacent frames in video. This poses a different kind of challenge to hard cut detection in video segmentation. Finally our wearable device captures images at a very low quality and uses a fisheye lens which accentuates the challenge even more so. Thus while it may appear on first examination that there are similarities between video segmentation and SenseCam image event detection, this domain poses different challenges to video structuring.

In terms of comparison with techniques applied to the traditional photo management domain, the techniques we are using and propose to use for measuring SenseCam image similarity do use the same low-level image similarity measures as can be used in comparing photos. However, there is no process applied in the traditional photo management domain which is equivalent to structuring SenseCam images into events. Images taken by existing digital cameras are manually taken at irregular intervals and events are normally determined by the time interval between successive images, as there are generally periods of hours, or even days or weeks, between successive images thus making it relatively straightforward to segment into events [64]. These techniques can't be used with wearable cameras as images are taken on average three times every minute. As with the video domain, it may initially appear that the lifelogging domain can take techniques from the traditional image domain to segment images into events. However, as we have just highlighted, the lifelogging domain presents very significant challenges on closer inspection.

In the lifelogging domain, Lin & Hauptmann propose a time-constrained variation of the *k*-means clustering algorithm to cluster images together based on their visual features while also considering what clusters their temporal neighbours belong to [89]. A weakness of this approach is that the number of events/clusters must be fixed to a value *k*, and also only one data source is considered (in fairness, Lin & Hauptmann only had video footage available [89]). Additionally they don't identify the boundaries between all events; instead of having multiple events of "working at PC" in a day, they will cluster all these events together. In this respect, the time-constrained clustering technique of Yeung & Yeo [157] used in the video domain for story boundary detection is more appropriate as it will retain all the local "working at PC" clusters (hence allowing us to identify the boundaries between them and their adjacent activities). Wang *et. al.* segment their lifelog of video into 5 minute clips, however real events are not always 5 minutes in duration [155]. In our published work we focused on investigating what individual and combined data sources yield the richest source of information for activity segmentation [50]. However, the selection of a threshold for the number of events in a day was arbitrarily fixed at 20, and only the precision of the segmentation was computed as we had no detailed groundtruth of valid event boundaries from which to calculate recall. In this published work we also discovered that recording audio is not necessary to segment lifelog data into distinct events/activities [50].

Adjacent groups of images are contrasted against each other to determine how different they are. Higher difference scores between images indicate a greater likelihood of this being a boundary between distinct lifelog events. In our work we compare adjacent sensor values, but we can also compare adjacent images by considering the pixel properties of each image. Images can be represented by characteristics such as the colours present in them, and also the edges present and the texture too. The MPEG-7 descriptor language is a standard introduced to recommend how images should be represented [125]. This standard details a range of descriptors for an image's colour (e.g. colour space, dominant colour(s), scalable colour, colour layout, colour structure, etc.), an image's texture (e.g. homogenous texture, edge histogram, etc.), an image's shape (e.g. region shape, contour shape, etc.), etc. In this thesis we use 4 low-level MPEG-7 descriptors: *colour layout* captures the spatial distribution of colour in an image, which is computationally efficient; *scalable colour* measures the colour distribution over an entire image, which is good at identifying general scenes (e.g. park is generally green, while workplace is another colour); while the *edge histogram* descriptor captures the spatial distribution of edges [112, 20], which is again good at identifying general scenes (e.g. not many edges when at a beach, but

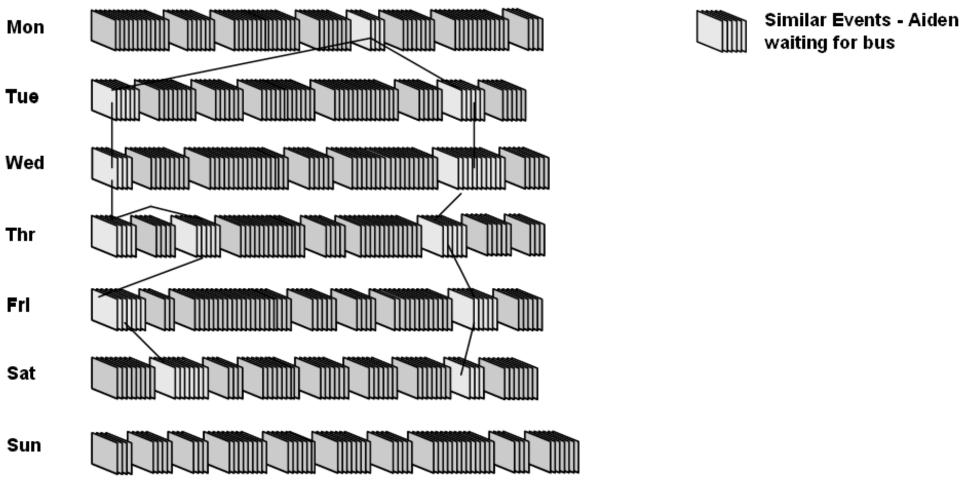


Figure 2.7: Retrieving other events that are semantically similar to a given lifelog event.

defined edges when working on PC); and *colour structure* measures the spatial structure of the colours in an image as well as their frequency of occurrence [100].

The MPEG-7 descriptors are represented by vectors, therefore it is necessary to compare the similarity of any given vectors [82, 124]. Also given a number of content and contextual sources are captured by the SenseCam it is also necessary to normalise [103, 132] and fuse [56] those sources together. This will be covered in more detail in Chapter 3 on event segmentation.

2.4 Retrieving Similar Events

Having detected the presence of distinct lifelog events, and given that many people in the lifelogging community have talked of the major challenge being that of *retrieval*, we now discuss how we approach retrieving other similar/related events to a given lifelog event (as illustrated in Figure 2.7).

To retrieve similar SenseCam events to a given SenseCam event in a lifelog it is necessary to firstly determine *how to represent* SenseCam events, and then *how to compare* those event representations against each other. We will now discuss each of these challenges in a separate subsection, and detail approaches taken by others in the literature.

2.4.1 Representing Events as Units for Retrieval

Representing an event (consisting of many images) can be done using a single image, analogous to a video keyframe. Smeaton & Browne [137] note that in video retrieval the middle frame is often chosen as the shot/event keyframe image [118, 51], which is then used as the unit to represent that entire shot. We investigate this method to represent lifelog events.

While selecting the middle image of an event is highly efficient in a computational sense, it may not always be the more representative image of the event. Cooper & Foote [41] investigate two keyframe selection methods, namely: 1) Selecting the individual image that is closest to all the other images in the event; and 2) Selecting the individual image that is closest to all the other images in the given event, but also is most distinct from all the other images in the other events. These video keyframe selection methods will be investigated to select images to represent events as units for retrieval.

Above we have described techniques that select a single image from the event to represent that event, however another approach is to take an average value from a selection of images in the event. In the video domain, FXPAL represent shots by the average of the block histogram values of all the images in each segment for the TRECVid 2007 summarisation task [34]. This technique is also investigated in this thesis as a unit for retrieval.

With regards to choosing a keyframe in the video domain, Smeaton notes that '*...The danger of choosing from the start or end of a shot is the increased likelihood of picking up artefacts from the previous shot if there has been a gradual rather than a hard shot transition ...*' [136]. This applies to the lifelogging domain too given the difficulty in achieving accurate event segmentation and as a result in this thesis we investigate if data near the middle of an event should be more strongly weighted than that at the start or end of an event.

2.4.2 Comparing Lifelog Events

Having decided how all the events in a lifelog should be represented, the next challenge is to determine *how to compare* given events to a query event from the lifelog.

Wang *et. al.* use the event of interest as a query video to find other similar videos based on visual and audio features [155]. We intend to extend the work of Wang *et. al.* [155] as instead of using 2 sources of information to find similar events, we display the benefits of using other contextual sources of information to find similar events.

Again given there are a number of content and contextual sources captured by the SenseCam it is necessary to normalise [103] and fuse [56] those sources together. The technical details of this will be covered later in this dissertation.

In an interesting contrast to the approaches described thus far, Hori & Aizawa ignore image content in retrieving related lifelog events because they found it would *probably* not be useful [72]. They retrieve similar events to a given lifelog event using contextual sources only; e.g. brain wave analyser, GPS, accelerometer and gyro sensors. However, in experiments detailed in Chapter 4 of this thesis we show that image content is very important in terms of retrieving relevant lifelog events to given query events captured by SenseCam users.

Voorhees provides an overview of how experiments are carried out in the text and video retrieval domains [154]. This thesis follows her experimentation methodology and also uses the same evaluation metrics: precision, precision@N, recall, F-Measure, and mean average precision. Precision refers to the percentage of similar events proposed by the system, which are relevant, while precision@N refers to the percentage of events in the top N items of a ranked list that are relevant. Recall refers to the percentage of similar events retrieved out of *all* the known similar events. To determine the recall a groundtruth or recall basis needs to be determined. The F-Measure is a trade off between both precision and recall $(\frac{2 * precision * recall}{precision + recall})$. Mean average precision is a metric to give an indication of how well a system returns many of the relevant results in highly ranked positions. For more details of these metrics, please refer to an overview paper written by Ellen Voorhees [154].

2.5 Determining Event Importance and Selecting Keyframe Images

Given that a user will on average capture over 7,000 events per year (approximately 20 per day), it will be useful to highlight those events that are more important. We envisage an interface showing keyframe images from each event, with more important events being given a large size/emphasis on screen as illustrated in Figure 2.8. In this section we discuss why such a facet is useful, and what approaches others have taken to attempt to automatically determine the importance of lifelog events.

In studies on the potential benefits of lifelogging to the human memory, Berry *et. al.* note that '*... Mrs B did not recall all of the autobiographical events recorded by SenseCam. For example, she remembered little of events three and nine. When discussing his wife's lack of recall for these events, Mr B said that they were "not memorable", in that they were "boring" to her. Mr B described them as "dreary days walking through dreary little towns ... Nothing stood out". Similarly, Mr and Mrs B did not think it necessary to record routine or mundane daily events ...*' [17]. Considering these comments we feel that it will be useful to automatically emphasise those lifelog events that may be of more interest to users, and to hide, or at least decrease the emphasis of, routine/mundane events.

In SenseCam user experiments Lee & Dey investigated the importance of selecting good images as cues for memory recall in which they found 4 general categories performed best as memory cues: people, the action taken, any object involved, and place [87]. They state the caregivers of people with Alzheimer's disease '*... experience great strain as they are required to support the increasing episodic memory needs in addition to the physical needs of normal aging ...*'. Firstly they shadowed caregivers in a home for a number of days in an ethnographic study. They found that the lives of the people with episodic memory impairments (EMI) are made more pleasurable when the caregiver gives them a cue to remembering a certain event, rather than just giving the EMI the whole story without giving them a chance to jog their memory. However, as the caregivers are under great strain,

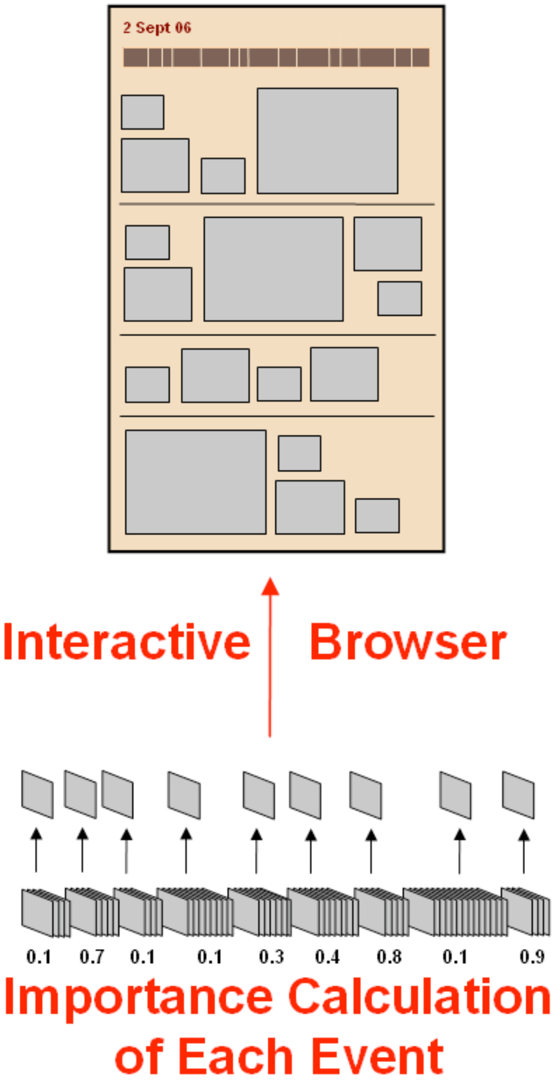


Figure 2.8: Identifying those events that may be more interesting to review and then displaying them appropriately to users.

sometimes they just blurt out the whole story in frustration. Considering this, Lee & Day then asked the people with EMIs to wear SenseCams, and afterwards printed out 100 photos of certain events. The researchers sat down with the EMIs and went through the photos to identify which individual images gave the best memory cues for a certain event. Eventually they narrowed each event down to the 6 images that provided the best memory cues. As stated at the beginning of this paragraph they found 4 general categories of cues (from 14 experiences) were most effective: people (7 of the 14), action (4), object (2), and place (1). Therefore to determine the importance of a given event or activity it may well be useful to highlight those events with people present. Determining the importance of events in the domain of lifelogging was first introduced by Tancharoen and Aizawa [145] and in their work they determine conversational scenes as being very important events. To detect these automatically, they used automatic face detection to determine events containing face-to-face conversations.

The concept of novelty detection has been used in the text [33] and video [58] domains. With respect to the lifelogging domain, novelty works on the premise that those events that are novel (e.g. rarely occurring events such as going to a rock concert) will be more important to a user than those events that occur frequently (e.g. having breakfast in the morning). The concept of novelty has not previously been used in relation to reviewing one's life in the domain of large collections of images captured by wearable cameras. Varian discusses ranking based on dissimilarity [152]: *'. . . Now, the interesting thing about the economist's notion of the value of information is that it is only new information that matters . . . it is better to look at risky choices early in the search process . . . a document that has a low expected payoff may still be presented earlier, if it has a chance of yielding a large payoff early in the search . . . in some cases, an optimal search involves looking at the riskiest items first!'*. Routine/mundane events are quite analogous to documents with a low expected payoff, while more dissimilar/novel lifelog events are quite analogous to the risky choices that we may decide to emphasise to the user. In the video domain Gaughan determines the most novel items/events being those that are visually most *dissimilar* to all the other items/events [57]. This thesis applies Gaughan's work to the lifelogging domain, and also extends it by incorporating contextual sources of information (i.e. from the sensors onboard the SenseCam).

Markou and Singh present an extensive review of various statistical approaches to detecting novelty in their journal paper [96], many of which are concentrated on defining correct threshold values. As this thesis uses novelty to rank events from most novel to least novel, it is not necessary to have a threshold to define whether an event is sufficiently novel or not.

We have been describing techniques to calculate the importance of events. Now the focus is shifted towards discussing methods to select a representative/keyframe image from each event (consisting on average of almost 100 images). Smeaton & Browne [137] note that in video retrieval the middle image is often chosen as the shot/event keyframe image e.g. [118, 51]. In literature this approach is generally used as the baseline on which to compare all other approaches. Other standard approaches include those discussed by Cooper & Foote [41] (see Section 2.4.1). In this thesis we will detail an approach to selecting a keyframe image by considering the "quality" of individual images, and then selecting the image in the event with the highest quality score. This will be described in more detail later in Chapter 5.

2.6 Event Augmentation

As described in the motivation scenario in the introductory chapter, many users may like to use lifelogging devices as a tourist tool, whereby they can passively capture images and thus enjoy the experience of being there without interruption. We envisage a scenario where the experience of users reviewing images of a special trip would be enhanced by having their lifelog event automatically augmented with images taken from external sources, as outlined in Figure 2.9. The novelty of our work is in augmenting SenseCam images with images/videos from external sources of data (i.e. the Internet), we now discuss somewhat

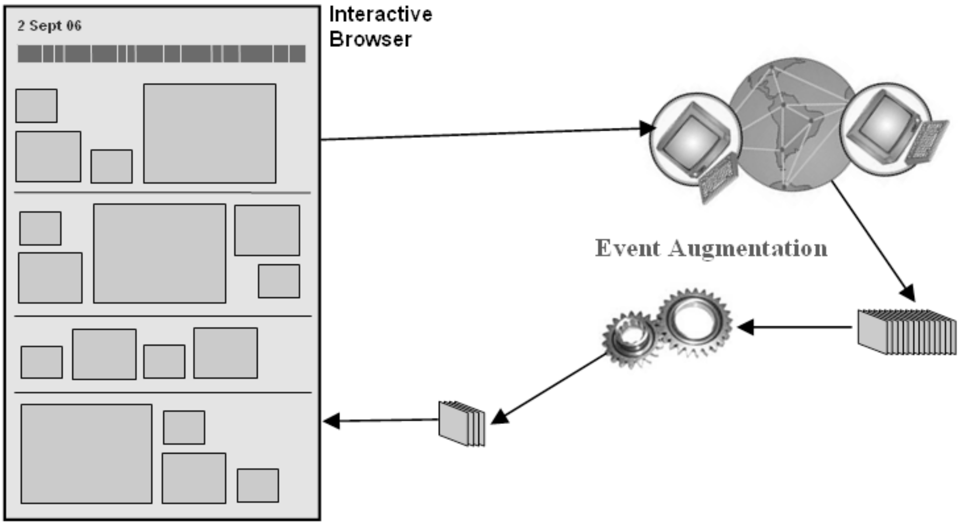


Figure 2.9: Augmenting lifelog events with images/videos from external sources of information.

related work carried out by others in literature.

The goal of the MyLifeBits project [60] is to store a significant amount of the experiences that an individual will encounter in his or her lifetime in digital format, and then to make that data searchable. Therefore the MyLifeBits project contains a database that logs many activities such as: Internet browser logging, interface activities, media files played, e-mails sent/received, SenseCam images, GPS location data, etc. This means that it is possible to augment SenseCam images with other pieces of information such as the location a picture was taken, whether the user was working with any particular application on their PC at the time a picture was taken, etc. It is our intention to augment SenseCam images with images from external data sources, and to investigate whether this is useful to users or not.

Lim *et. al.* attempt to recognise images taken by tourists on mobile phones, and then augment the tourist's photo with a textual description of a recognised monument [88]. They call this system the Snap2Tell system, whereby they have the STOIC database which contains 5,278 images of 101 Singapore tourist locations (they have a minimum of 16 images per scene). They then do a novel form of matching image patches around salient regions of the images, and return the database image closest to the picture taken by the tourist on their mobile phone. The image is returned along with a textual description of the attraction to the phone. A weakness of this approach is that the database of images is restricted to 5,278 images. O'Hare *et. al.* have done something very similar too [115]. In essence they selected 6 example query images and attempted to identify similar photos from other users' collections. They filtered firstly based on location, investigating the optimal distance to filter results by. There didn't appear to be a significant difference in filtering by 200m (smallest distance) or by within the same city (around 5-10km), probably due to the small dataset available. They did an image:image similarity matching between the query image and the resultant images. The main conclusion of their work is that it is necessary to firstly filter by location before attempting matching images based on low-level features. Nini & Batouche have done similar work in that they try to identify fixed objects and then provide extra augmented information on those objects [110]. Like the work of Lim *et. al.* [88] a limitation of all these systems is that the dataset of images is of a fixed size. Kawamura *et. al.* also produce a memory augmentation paper, but it is based on using RFID tags to identify objects in a *fixed* scene [76].

Since we expect lifelogging devices to become more commonplace in future, and given the phenomenal growth of multimedia content on websites, we introduce the idea of augmenting lifelog events with "Web 2.0" content, where individual users make small contributions in uploading images/videos. We investigate aggregating these small contributions over an enormous scale of users, which can thus automatically enrich the experience of individuals reviewing their trips or events, by providing them with a large number of relevant items of information mined from millions of other individuals.

Kennedy and Naaman detail a clustering method to select images of a location that are both representative and also distinct, e.g. showing images from the Flickr "Web 2.0" site taken from multiple view points of Big Ben in London, and also pictures of other landmarks in London too [80]. They focus on diversity of results, and in effect the information task is displaying images on a browser. The information need of our event augmentation work is different as it's very specific, where users only want to see images of say the Eiffel Tower,

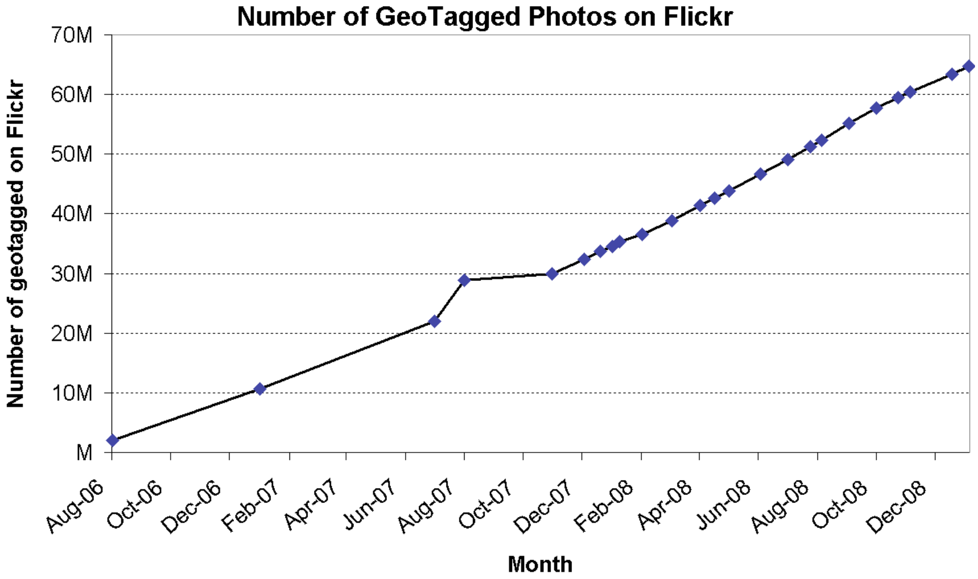


Figure 2.10: This graph illustrates the phenomenal growth of geotagged images on the Flickr website.

and not other regions in Paris.

The work in this thesis will attempt to find similar pictures in the same location as those lifelog images taken by a wearable camera using the Flickr[6] and Panoramio[7] websites. The Flickr site has over 65 million geo-tagged images (at time of writing) and is growing at a rate of approximately 500,000 geotagged images per week as illustrated in Figure 2.10. This overcomes the limitations of other systems which have only been able to augment lifelog images with limited datasets of information [88, 115, 110, 76]. The significant research challenge here is to find how to identify the relevant images from the Flickr website to a given lifelog event. Afterwards, we also explore augmenting lifelogs with images from other non-geotagged sources on the Internet such as Yahoo!, MSN, and YouTube; this will be covered in much greater detail in Chapter 6 of this thesis.

[6] www.flickr.com
[7] www.panoramio.com

2.7 Summary

Initially in this chapter we have pointed towards evidence that visual images are an excellent memory stimulant, and that in particular humans can retrieve past memories more easily when they are provided with visual cues towards those memories. We pointed out that many people have memory impairments and that technology perhaps may be useful as a memory aid, particularly given that cued recall is better at retrieving autobiographical memories than free recall. This motivated us to then discuss the history of lifelogging and that in the past people kept diaries to record mundane events. However, this required much effort and even as far back as 1945, Vannevar Bush recognised that someday in the future we would have the capability to automatically record many aspects of our lives [28]. As computing technology advanced from the late 1970's onwards, so did the efforts in creating and miniaturising lifelogging devices, with Steve Mann being at the forefront of such efforts [94]. The big challenge now facing the lifelogging community is the retrieval of information that is relevant to end users, and this has been noted by many respected researchers in the community [61, 60, 3, 144]. We have also discussed a study carried out by researchers from the memory clinic in Addenbrooke's hospital in Cambridge, U.K., which indicates that lifelogging applications may indeed be beneficial memory for people with memory impairments [17].

The work in this thesis is concentrated on the visual lifelogging domain and in particular on the SenseCam which is the prevalent device in this domain, especially with respect to memory-based experiments, i.e. we concentrate on how to automatically structure and manage images generated by the SenseCam. We have motivated that the important challenges include being able to:

- effectively segment a large collection of personal images into distinct events
- accurately determine which of those events are most unusual
- usefully provide the ability to link similar events together

- augment images from the low-resolution wearable device with higher quality images from external data sources.

In this chapter we have discussed the work carried out in literature by other research groups attempting to address these challenges. With respect to segmenting sequences of images into distinct events or activities, we have shown that past research was conducted on single sources of information and tested on small datasets. In determining the most important events in a person's lifelog, face detection has been used to identify events containing face-to-face social interaction. We also discuss novelty detection techniques used in other domains, and why we investigate it in the work towards this thesis. In terms of retrieving similar SenseCam events to a given event, others have used either image content only, or sensory context only; in this thesis we will examine the combination of both together. Finally we introduce the novel concept of augmenting one's lifelog with images from the Internet and "Web 2.0" websites.

The next chapter of this thesis will discuss our approach to segmenting sequences of SenseCam images into distinct events or activities. This chapter is the underlying basis for the work discussed in the other chapters that succeed it, e.g. before we can search for similar activities to a given activity, we firstly must identify *all* the activities in the entire dataset. After discussing contributions made towards the segmentation of lifelog images into distinct events/activities, we will then discuss retrieving similar events to a given SenseCam event (Chapter 4), determining keyframe images for events and how important they are (Chapter 5), augmenting events with images from external sources (Chapter 6), and small-scale experiments to investigate the usefulness of these approaches as cues to retrieving autobiographical memories in a system presented to users (Chapter 7).

Chapter 3

SEGMENTING SEQUENCES OF IMAGES INTO EVENTS

> *In order for visual lifelogging technologies to effectively provide autobiographical memory retrieval cues it is necessary to automatically:*
>
> . . .
>
> - *effectively segment a large collection of personal images into distinct events*
>
> . . .

A SenseCam captures 2,000 images on an average day creating a sizable collection of images even within a short period of time, e.g. over 700,000 images per year. To help manage such a substantial quantity of information it is important to automatically split these collections into manageable segments (Figure 3.1) by identifying the boundaries between different daily events, e.g. having breakfast, working in front of a computer, attending a game of football, etc. (Figure 3.2). This takes advantage of the fact that *"...segmenting ongoing activity into events is important for later memory of those activities ... "* [158]. In this dissertation we will chunk images into semantic events, and this chapter will now provide details on how to effectively achieve this aim.

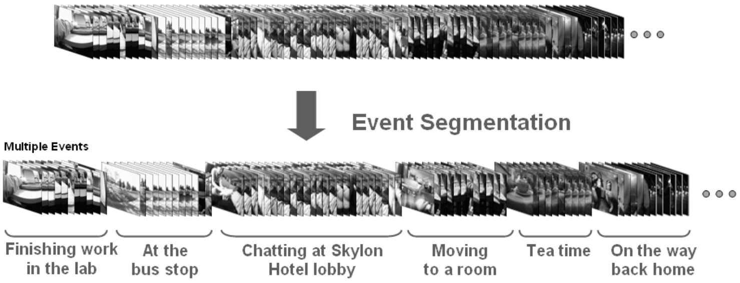


Figure 3.1: Segmenting a day's of SenseCam images into distinct events.

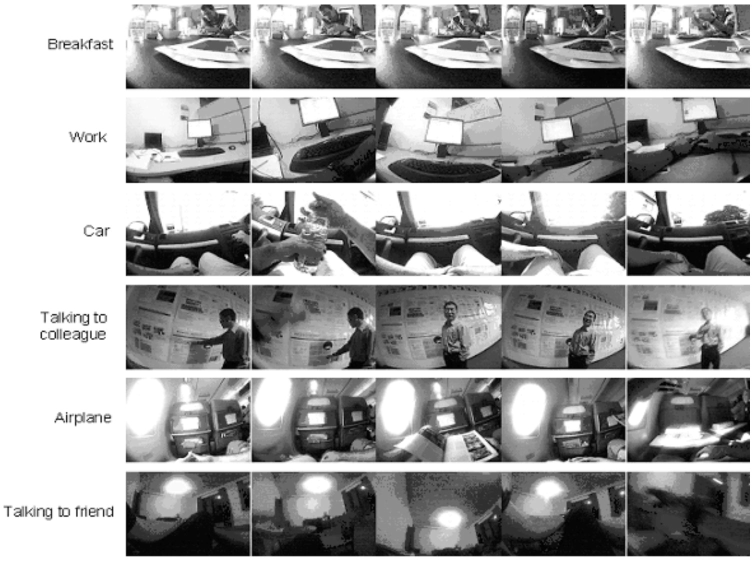


Figure 3.2: An example of distinct activities. The aim of this chapter is to identify the transition between events/activities.

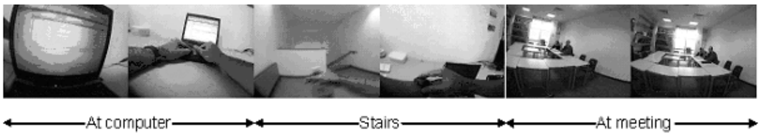


Figure 3.3: An example of a boundary between activities.

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, or was watching TV and then goes to prepare a meal, we believe it will be 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 3.3. In essence the aim is to detect moments of *change*, whether they be visual, sensory, or otherwise.

This chapter investigates the performance of numerous event segmentation techniques.

3.1 Processing of Images into Events

In this section details will be provided on our algorithm which segments sequences of lifelog images into distinct events/activities. The challenge of segmentation is very important, and one on which the other chapters of this thesis rely upon.

In our approaches to segmentation sequences of SenseCam images are firstly broken up into a series of chunks, where the boundary between these chunks correspond to periods when the device has been turned off for at least 2 hours (e.g. when the user has gone to sleep). Usually each chunk corresponds to a day's worth of SenseCam images.

Each image is then represented by image and sensor values. Use was made of the ace-Toolbox [112], a content-based analysis toolkit based on the MPEG-7 eXperimental Model (XM) [125], to extract low-level image features for each image. There are a multitude of available MPEG-7 descriptors, and we have proposed 4 that are particularly well suited towards SenseCam images. To recap, the MPEG-7 descriptors we use are colour layout, colour structure, scalable colour, and edge histogram [20]. The colour layout descriptor is to capture the spatial distribution of colour in an image; the colour structure descriptor measures the spatial structure of the colours in an image as well as their frequency of occurrence [100]; the scalable colour descriptor measures the colour distribution over an entire image; while the edge histogram descriptor captures the spatial distribution of edges [112]. It takes approximately 30 minutes to process a larger than average day of 2,500 images on a 2.4GHz Pentium 4 machine with 512Mb RAM. These MPEG-7 descriptors are represented as vector values. The sensor values associated with each and every image include: accelerometer (tri-axis)[1], ambient temperature, light level, and passive infrared detector.

To segment a day of images into distinct events, as outlined in Figure 3.4, processing follows these steps:

- Compare adjacent image/sensor values against each other to determine how dissimilar they are (Item 2 in Figure 3.4).
- Combine the various data sources together in an optimal manner (3).
- Determine a threshold value whereby higher dissimilarity values indicate areas that are likely to be event boundaries (4).
- Remove successive event boundaries that occur too close to each other (5).

3.1.1 Comparing Adjacent Image/Sensor Values

The motivation for comparing adjacent images against each other is to identify periods of time in which there is a large dissimilarity in visual or sensory values, thus indicating a high probability of change in activities undertaken by the wearer. To achieve this aim there are 2 main areas of investigation:

- How to quantify the difference between adjacent image/sensor values?
- How to select values for comparison i.e. is it optimal to compare single adjacent values or blocks of adjacent values?

[1] The 3 units are combined together in a Euclidean type fashion as detailed by Ó Conaire *et. al.* [38]

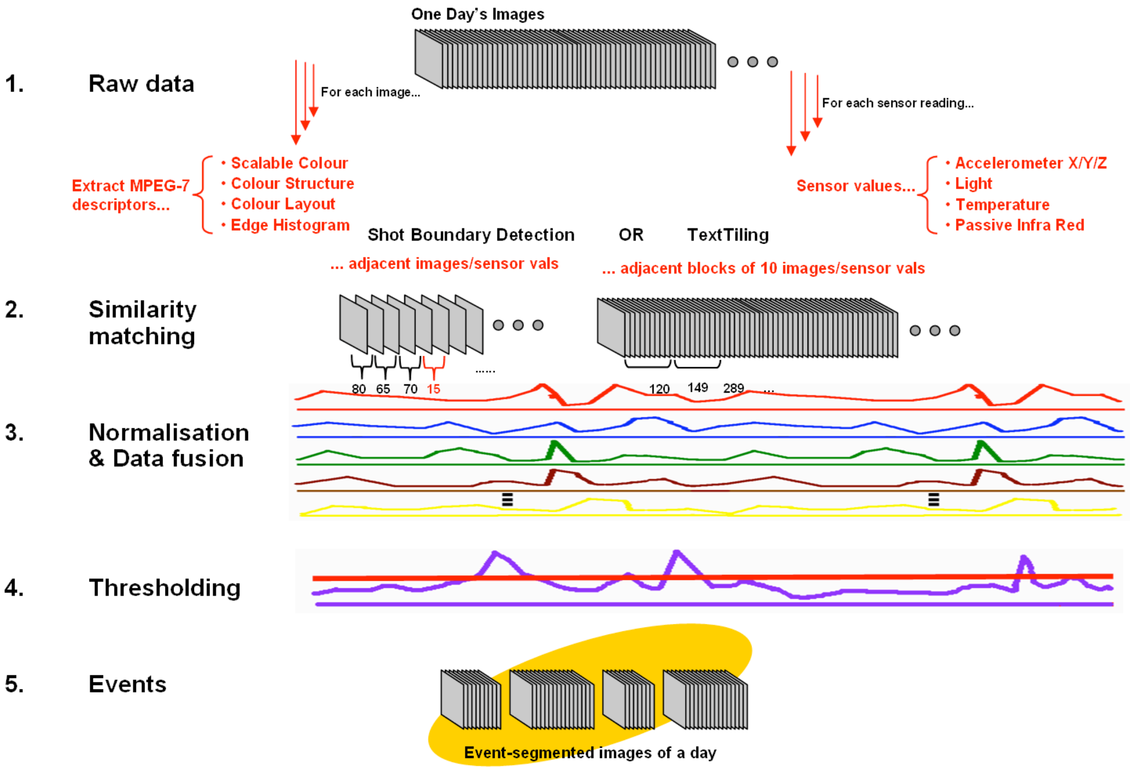


Figure 3.4: Overview of how we automatically segment sequences of SenseCam images into distinct events.

3.1.1.1 Quantifying Difference Between Adjacent Image/Sensor Values

As the sensor sources of information are all represented by single scalar values it is straightforward to compare sensor readings from adjacent readings. To calculate the difference between two light level readings, x and y, the answer is d_s in Equation 3.1 below:

$$d_s(x, y) = |x - y| \tag{3.1}$$

However the image MPEG-7 features are represented by vector values. The colour layout feature is represented by a vector of 12 bin values; colour structure is a 32 bin vector; scalable colour is a 64 bin vector; and edge histogram is an 80 bin vector. These values are concatenated together to produce a 188 bin vector i.e. early fusion. However, to calculate the difference/dissimilarity between two adjacent images (represented by MPEG-7 vector values), there are numerous approaches that can be taken. In experiments carried out in this thesis a number of vector distance/similarity techniques were investigated, particularly since no extensive experiments have been carried out to investigate these in the domain of lifelogging before. The following metrics were investigated:

Euclidean Distance: Discussed in [82]

$$d_{euc}(x, y) = \sqrt{\sum_{i=1}^{d} (x_i - y_i)^2} \tag{3.2}$$

Manhattan Distance: Discussed in [82]

$$d_{man}(x, y) = \sum_{i=1}^{d} |x_i - y_i| \tag{3.3}$$

Squared Chord Distance: Discussed in [82]

$$d_{crd} = \sum_{i=1}^{d} (\sqrt{x_i} - \sqrt{y_i})^2 \tag{3.4}$$

Square Chi-Squared Distance: Discussed in [82]

$$d_{chi}(x, y) = \sum_{i=1}^{d} \frac{(x_i - y_i)^2}{x_i + y_i} \tag{3.5}$$

Canberra Distance: Discussed in [82]

$$d_{can}(x, y) = \sum_{i=1}^{d} \frac{|x_i - y_i|}{|x_i| + |y_i|} \tag{3.6}$$

Bray-Curtis Distance: Discussed in [82]

$$d_{bry}(x, y) = \sum_{i=1}^{d} \frac{|x_i - y_i|}{x_i + y_i} \tag{3.7}$$

Histogram Intersection Distance: Discussed in [124]

$$d_{his}(x, y) = 1 - \frac{\sum_{i=1}^{d} min(x_i, y_i)}{\sum_{i=1}^{d} y_i} \tag{3.8}$$

Kullback-Leibler Distance: Discussed in [124]

$$d_{kul}(x, y) = \sum_{i=1}^{d} x_i \log \frac{x_i}{y_i} \tag{3.9}$$

Jeffrey modification of Kullback-Leibler Distance: Discussed in [124]

$$d_{jef}(x, y) = \sum_{i=1}^{d} (x_i \log \frac{x_i}{m_i} + y_i \log \frac{y_i}{m_i}) \tag{3.10}$$

where $m_i = \frac{x_i + y_i}{2}$

X^2 **Statistics Distance**: Discussed in [124]

$$d_{xst} = \sum_{i=1}^{d} \frac{(x_i - m_i)^2}{m_i} \tag{3.11}$$

where $m_i = \frac{x_i + y_i}{2}$

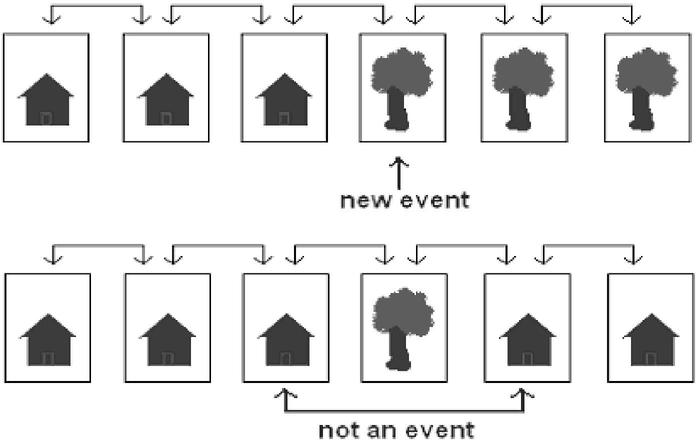


Figure 3.5: Illustration of possible false positive event boundaries.

3.1.1.2 Selecting Number of Adjacent Values to Compare

Having presented a number of techniques for comparing individual images against each other to see how similar/different they are, we now discuss how to select values for comparison i.e. is it optimal to compare single adjacent values or blocks of adjacent values? The result of this processing stage is, for the image and sensor data sources, is a list of similarity scores. Those instances of adjacent images/sensor values being more different indicate times where it is more likely that the user has began a new activity.

As just stated, if adjacent images are sufficiently dissimilar 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 captures from the perspective of the wearer. Also an average of 22 seconds exists between adjacent captured images. 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 only adjacent images are compared the wearer looking momentarily in the opposite direction would trigger an event boundary, as both images are very different in their visual nature. This phenomenon is illustrated in Figure 3.5.

We address this problem by using an adaptation of Hearst's Text Tiling algorithm [70]. This involves the comparison of two adjacent blocks of images against each other, to deter-

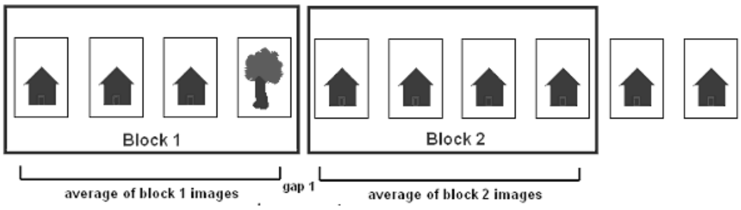


Figure 3.6: Our adaptation of TextTiling.

mine how similar they are. As an example in Figure 3.6 in 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 (using the average vector value of all the images in each block) then it is quite likely no event boundary has occurred, however if the two blocks are sufficiently dissimilar, 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 3.6 with the house and tree icons.

3.1.2 Combining Various Data Sources

After comparing adjacent image and sensor values against each other, there will be a separate list of *difference values* for each individual source. The greater the difference value, the greater the likelihood that an event boundary has taken place. However, sometimes a flash of bright light may make the light sensor indicate that an event boundary has taken place, while in fact the other sensors have recorded no notable change taking place. Therefore we investigate the benefits of fusing together all available sources of information, and we now detail our approach in achieving this.

Montague and Aslam note that *"A successful* [search] *system will consistently improve on the best of its inputs, no matter how many input systems are available"* [103]. Given that there are numerous sources of information available in the lifelogging domain the challenge of combining these sources optimally is quite important. The SenseCam device presents 5 sources of information: the content of the actual image, the accelerometer/motion sensor, the ambient temperature, the light level sensor, and whether the passive infrared sensor detects body heat or not. The process of combining multiple sources has two distinct stages, firstly normalisation, and then fusion.

There are two different types of fusion, early fusion and late fusion. Early fusion refers to when sources of information are combined together before they are manipulated to indicate the resulting output, whereas late fusion refers to when sources are firstly manipulated to produce an output and are then combined together. In the segmentation work in this thesis the 4 MPEG-7 descriptors are early fused together to produce a 188 bin vector, as detailed in the previous section. All other fusion discussed in this section, from this point onwards, is late fusion i.e. where the sources have already been manipulated so that they provide an indication of an image being an event boundary. Then those scores can be fused.

3.1.2.1 Normalisation of Data Sources

Before data sources can be combined together it must be ensured that they are all on the same scale, e.g. from a sample day of 2,006 images (taken by one user on 30/03/07) the difference in the light level for successive images returned a scalar value of up to 9,587 (with an average of 982), whereas the difference in the ambient temperature reading of successive images only returned a scalar difference as large as 13 (with an average of 0.23). Consequently it is important to normalise the readings from all the data sources to a common scale, so that they can be compared to each other in a fair manner, and fused correctly. A number of techniques proposed by Montague and Aslam [103] are investigated in this work:

Min-Max Normalisation: Shift all scores so that the minimum score is zero, and then scale the resulting values so that the maximum score is equal to one (i.e. divide the resulting values by the maximum score).

Sum Normalisation: Shift all values so that the minimum score is zero again, but this time scale the resulting values so that their sum equals one (i.e. divide the resulting values by the sum of all their scores).

Mean-Shift Normalisation: Shift all scores so that the mean of those scores is zero, and then scale the resulting values so that the standard deviation is equal to one (i.e. divide the resulting values by the standard deviation of all their scores).

Also a fourth technique was investigated, which is an adaptation of the Min-Max normalisation, **Standard Normalisation**: Scores are not shifted, but are simply scaled by the maximum score (i.e. just divide all the raw values by the maximum score)

3.1.2.2 Fusion of Data Sources

Once the data sources have been normalised to a common scale, the process of combination, or fusion, can be carried out. Again as with the normalisation stage, no experiments have previously been carried out in the lifelogging domain to indicate the optimal information fusion approaches. Fox and Shaw propose a number of fusion techniques [56] that are applied to the lifelogging domain in this thesis:

CombSUM: Resultant value is the sum value of all the sources

CombMAX: Resultant value is the maximum value from the source

CombMED: This time the median value is selected

CombMIN: The minimum value is selected

CombANZ: Return the sum of values *divided* by number of non-zero scoring sources

CombMNZ: Return sum of values *multiplied* by number of non-zero scoring sources

As mentioned, there are 5 data sources, each of which may be used, to indicate how likely an image is to be an event boundary i.e. the larger the difference in adjacent values, the more likely it is to be an event boundary. However, it is possible that not all 5 sources are good at identifying event boundaries, therefore numerous combinations of sources will also be tested. Also particular sources may be very good at identifying event boundaries, therefore it is desirable to attach a greater confidence, or weighting, to such sources of information. Consequently the confidence/weighting to be assigned to each source of information is also investigated in this thesis.

3.1.3 Peak Scoring and Determining Threshold to Identify Event Boundaries

The previous processing stages produce a list of scores indicating how likely each is to be an event boundary. The next stage involves selecting the actual instances that are most likely to be event boundaries. This section covers two approaches to automatically realise this stage of processing. Firstly the benefits of exaggerating, or emphasising, the most likely images are investigated, and secondly the selection of an appropriate threshold is explored.

3.1.3.1 Peak Scoring

Given a list of scores, as illustrated in Figure 3.7, the aim is to have them match the user groundtruth boundaries (the vertical lines). Higher scores indicate a greater likelihood of an event boundary taking place. In this case it is desirable to exaggerate the instances where peaks occur. To achieve this we introduce a method referred to as *peak scoring*. Each individual score is compared to its adjacent neighbours, for as long as they are monotonically decreasing. The individual score in question is then changed to equal the value of its difference to the smallest leftmost neighbour added to the difference between it and the smallest rightmost neighbour (where the scores have also been monotonically decreasing). This is illustrated in Figure 3.8 whereby the highlighted instance will have a new score of $h1 + h2$. We refer to this approach as *peak scoring* for the remainder of this thesis. The bottom half of Figure 3.8 illustrates the effects of the peak scoring approach on the values of Figure 3.7. The result is a more “spiky” graph where images with higher scores are exaggerated much more sharply than other “noisy” images (with respect to their likelihood of being an event boundary). This thesis investigates if the proposed technique is beneficial towards the automatic segmentation of images into events.

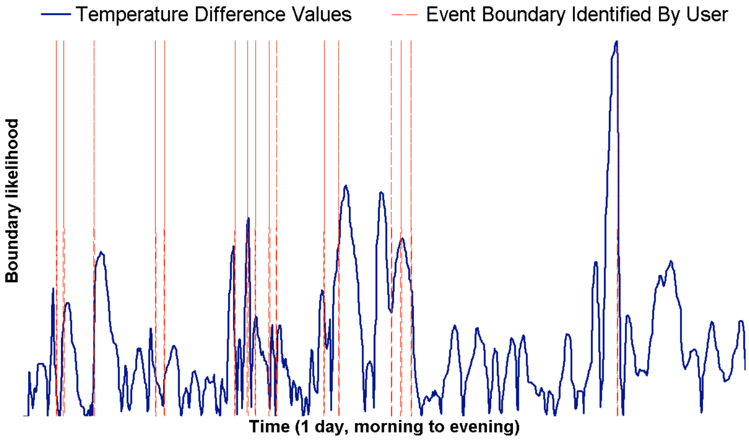


Figure 3.7: Event boundary likelihood scores from a sample day using the temperature source of information before peak scoring.

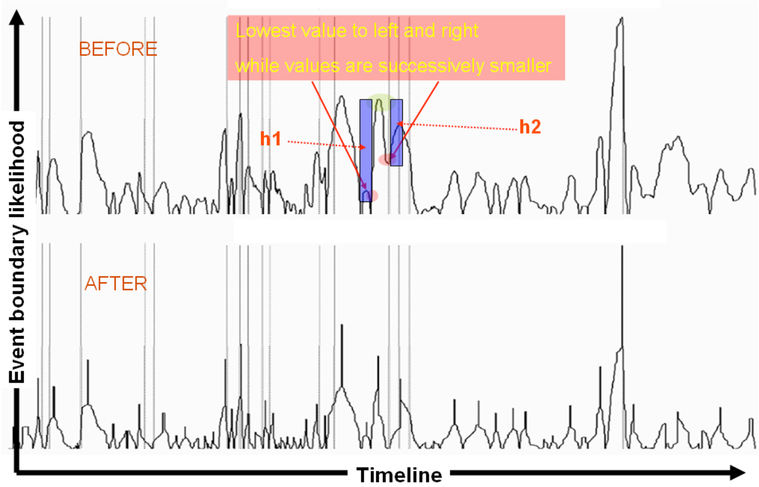


Figure 3.8: An overview of how our peak scoring approach functions. For a given peak, we consider the difference to the lowest value to its left (*h1*) while scores are monotonically decreasing, and similarly the difference to lowest value to its right (*h2*). These two difference scores are then added to give a new *peak score*, which is a more “spiky” value as can be viewed in the bottom graph.

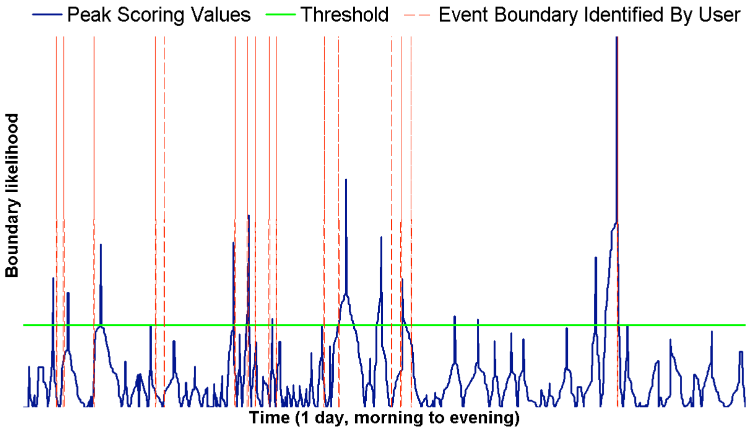


Figure 3.9: All values above the line in question are selected as the event boundaries for this sample day - this highlights the importance of selecting an appropriate threshold value.

3.1.3.2 Threshold Selection

All the previous stages gave a likelihood of each instance being an event boundary between sequences of images, however no decision was made on *which* instances should be selected as the final event boundaries. Therefore, following peak scoring our next task is to automatically determine which instances should be chosen as boundaries (Figure 3.9). We do this by automatically choosing a threshold value based on all the peak scoring values just produced. All values above the threshold line are selected. As a result if the threshold value is selected too low, there will be a number of false boundaries detected; however if the threshold is too high a value, there will be a number of valid boundaries undetected. In this thesis we investigate two thresholding techniques, one non-parametric (Kapur) and one parametric (Mean).

Kapur thresholding [75] is an entropy based technique that *automatically* selects the relevant threshold given a range of values. All the values are firstly divided up into a number of histogram bins e.g. if range of values was from 1-6, and 2 bins are used, then bin 1 will contain the count of elements that are in the range 1-3, and bin 2 will contain the count of elements that are in the range 4-6. The Kapur method then automatically selects a threshold dividing the histogram into two probability distributions, one representing the object/event and one representing the background noise [37]. Work in this thesis investigates the optimal number of bins to select for Kapur thresholding in the lifelogging domain. The threshold is set to the value of bin *j* that maximises E in Equation 3.12.

$$E = \frac{E_j}{A_j} + logA_j + \frac{E_N - E_j}{A_N - A_j} + log(A_N - A_j) \tag{3.12}$$

Mean thresholding [133] is a parametric technique that selects the relevant threshold given a range of values. This method approximates that the background noise is Gaussian and is therefore the dominant mode in the histogram. Consequently the threshold is selected by adding *k* standard deviations to the mean [37]:

$$T_{mean} = \mu + k\sigma \tag{3.13}$$

Work in this thesis investigates the optimal value of *k* with regards to the lifelogging domain.

3.1.4 Post-Processing: Boundary Gaps

At certain times, such as running to catch a bus, there may be a large amount of visual and sensory change in successive images, and neighbouring images may be proposed as discrete events, even after thresholding. In reality only one of these should be selected as the event boundary. In effect we are seeking to explore the minimum length of time that every event must be. Therefore we investigate, for an image proposed as a boundary, the optimum amount of time to ignore subsequent images proposed as boundaries, e.g. 2, 3, 4, 5, 10, etc. minutes, as illustrated in Figure 3.10. We refer to this approach as *post-processing boundary gap* in the remainder of this thesis.

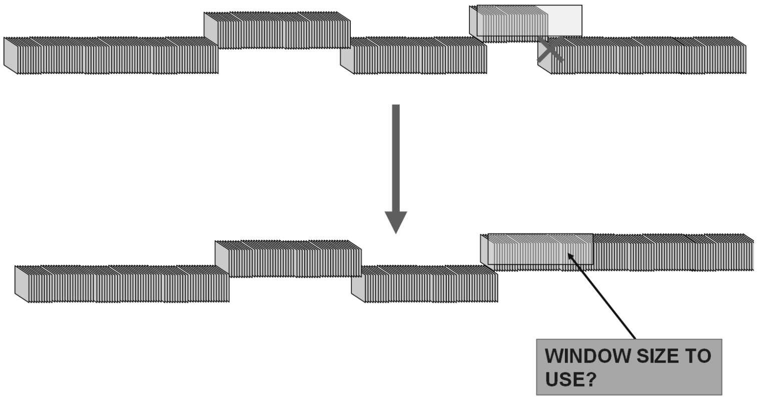


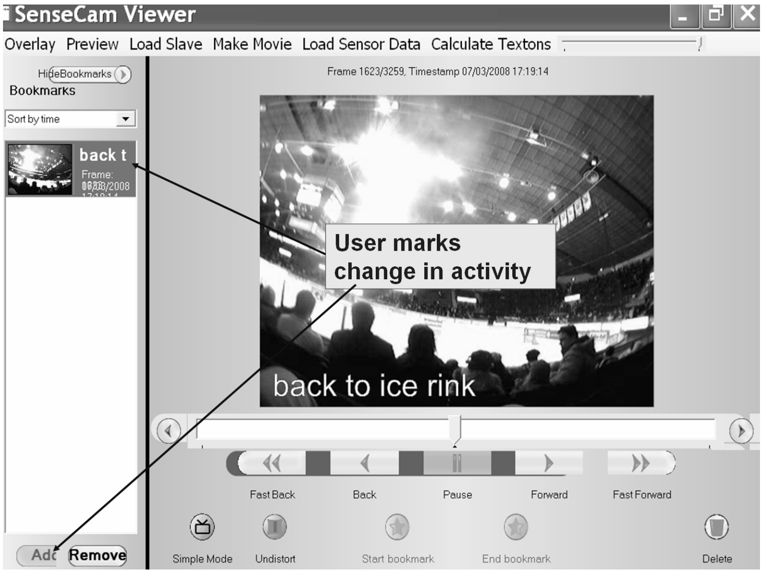
Figure 3.10: An overview of our post-processing boundary gap method.

3.2 Event Detection Experimental Setup

We now describe the setup for our event detection experiments on 5 users (information retrieval experts aged 25-35) who each wore a SenseCam over a one month period. Afterwards each user reviewed their SenseCam collection and manually marked the boundary image between all events for each day, to create a groundtruth. It is important for privacy reasons that the owner of the lifelog images is the person who groundtruths those images, as by their nature lifelog images are highly personal.

Users made use of the Microsoft SenseCam Image Viewer software [71] to identify images that were at the boundary of two distinct events/activities (Figure 3.11). Users could create a bookmark on the left panel of images they believe to be boundaries between two semantically distinct events. The main panel of the application is for playing all of the lifelog images as a video, with buttons to play, pause, fast forward, rewind, etc.

Kahneman *et. al.* detail the activities of 909 women where they carried out a questionnaire, and prompted the women to list down their activities and how they felt during these activities [74] . They discovered that there was an average of 14.1 episodes during the day and that the average episode duration was 61 minutes. We believe that more events would have been identified if the users had access to digital lifelog data of their days, as they would have improved recall. In the experiments for this section of the thesis users



User marks change in activity

Figure 3.11: Microsoft Research SenseCam Image Viewer Application [71].

User	Avg Daily Duration	Total Num Images	Groundtruthed Events	Avg Events Per Day	Avg Images Per Day	Images Per Event
1	13h 08m	80,934	995	28	2,312	81
2	9h 27m	76,810	875	18	1,600	88
3	10h 41m	44,447	348	17	2,116	128
4	7h 45m	27,929	329	13	1,117	85
5	9h 15m	41,043	439	19	1,783	93

Table 3.1: Statistics on data gathered by our users for event segmentation experiments.

were asked to manually identify event boundaries, and it was stressed to these users to judge event boundaries based on the *semantic* meaning for that user personally. Table 3.1 provides a breakdown of the 271,163 images captured and groundtruthed by the 5 users. It is interesting to note that here users identified an average of 19.1 events/activities/episodes per day, which somewhat corresponds to the 14.1 that Kahneman *et. al.* found in their study, when taken into account that the 5 users in the study in this thesis were reviewing a digital lifelog (which aids recall [13]).

It is interesting to note that in total there were 2,986 boundaries annotated in our set which represents a sizeable collection. This data was split into a training and test set. As mentioned at the beginning of the previous section, sequences of SenseCam images are broken up into a series of chunks where each chunk usually corresponds to a day's worth of SenseCam images. Not wanting to segment these chunks when dividing the data into training and test sets, the data was divided as closely as possible to give training and test set sizes of 50%. This results in a training set of 139,097 images (51.3%) and a test set of 132,051 images (48.7%, corresponding to 62 day's worth of images). The complete set of test images can be segmented in less than 2 minutes of CPU time on a desktop machine and over 3,000 parameter combinations were tested and measured against the groundtruth. Some of these approaches will now be presented in the next section.

3.3 Evaluation/Results

We now discuss the effectiveness of our approach to the detection of boundaries between events. We evaluate numerous variations of our own approach, and eventually compare to prior techniques used in segmenting sequences of lifelog images into events. To evaluate the effectiveness of each approach, three different metrics were used: precision, recall, and F1-Measure:

Precision: This is a measure of the percentage of boundaries proposed by the system that are accurate;

Recall: This is a percentage measure of how many of all the groundtruth boundaries were identified;

F1-Measure: This is a trade off between recall and precision, $F1 = \dfrac{2 * precision * recall}{precision + recall}$

.

Unless stated, results in this section are reported in terms of F1-Measure as it is desired to maximise both precision and recall. Parameters were optimised on the training set, and all results reported are on the test set.

Vector Distance Method	F1-Measure
Histogram Intersection	0.6271
Euclidean	0.6253
Manhattan	0.6166
Squared Chord	0.6023
Jeffrey Mod KL	0.6020
Bray Curtis	0.6013
Square Chi Squared	0.5907
X^2 Statistics	0.5905
Kullback Leibler	0.5869
Canberra	0.5684

Table 3.2: Segmentation performance of vector distance methods.

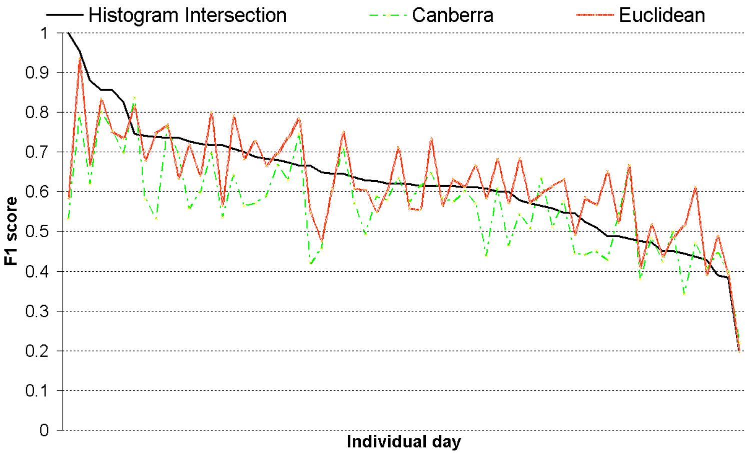


Figure 3.12: Breakdown of vector distance performances by day.

3.3.1 Best Vector Distance Method

As images are represented by MPEG-7 vector values, a number of vector distance/similarity techniques were investigated, and it should be noted that no extensive experiments have been carried out to investigate vector distance techniques in the domain of lifelogging before. Table 3.2 shows the performances of the 10 similarity measures for image vector comparison which we investigated, and shows the *Histogram Intersection* distance (which was also the best trained parameter) performs best overall.

As is be observed in Figure 3.12 the *Histogram Intersection* approach performs better than the *Canberra* approach (which was the worst performing overall) most of the time (on

44 out of 62 days), but not always (*Canberra* better on 16 out of 62 days). When comparing *Histogram Intersection* approach another high performing approach like *Euclidean*, there is no clear best method, as *Histogram Intersection* is better on 31 days while the *Euclidean* approach is better on 30 days.

3.3.2 Effects of TextTiling

As explained earlier, due to the nature of the SenseCam taking 3 snapshot images per minute, there may be certain images where the user briefly looked or moved in an opposite direction. While semantically this may be meaningless, it may be recorded as significant by one or more of the image and sensor sources of information. Therefore we introduced the *TextTiling* technique whereby blocks of values are compared against each other, so as to smooth out any outlier values.

The TextTiling approach was found to perform better on average (than non-TextTiling) for the Image (0.6023 vs. 0.5387), passive infrared (0.5151 vs. 0.0844), and temperature (0.4854 vs. 0.4218) data sources[2]. The optimal TextTiling block size for the image source was to use the average value of 5 images grouped together, while for the temperature and passive infrared sensors it was optimal to use a block size of 8 images. The latter are sources of information that change more slowly, e.g. the ambient temperature value will change slowly over time.

However comparing individual adjacent values performs better than TextTiling for the sources of information that do change quickly (e.g. motion values change quickly when user is sitting down and then decides to walk to another location) namely the accelerometer (0.5284 vs. 0.3307) and light (0.5209 vs. 0.3988) sensors.

Table 3.3 summarises the usefulness of TextTiling *overall on average* for each source.

On inspection of Figure 3.13 it is noticeable that initially recall is better than precision, but recall drops off at quite a fast rate as the window size increases, while the precision

[2] The performance figures calculated here are not final system results as some of these sources were later omitted in the fusion stage. Therefore in this subsection results reported are just the overall segmentation performance using each source individually, hence the performance isn't against the baseline F-Measure of 0.6271 reported in all other subsections

Data Source	TextTiling Optimal?	Block Size (num images)
MPEG-7 Image	TRUE	5
Accelerometer	FALSE	1
Light	FALSE	1
Temperature	TRUE	8
Passive Infrared	TRUE	8

Table 3.3: This table lists the information sources where TextTiling is useful.

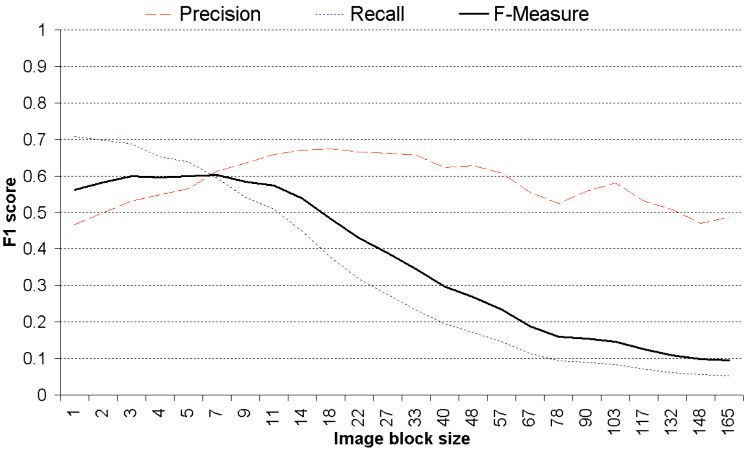


Figure 3.13: Note the effects of changing the TextTiling block size on the image source of information.

initially increases, before gently decreasing. This is to be expected and holds true for all the other sources, as by taking larger block sizes, more significant events will be detected (e.g. large visual change between being in car and then in front of PC), however fine-grained events will be missed (i.e. the event may have a shorter number of images than the size of the TextTiling block).

Finally Figure 3.14 illustrates the performance of the TextTiling technique compared against the Shot Boundary Detection technique on the MPEG-7 image source for each day in the test set from all the users. It is evident that while the TextTiling approach works better overall (0.6023 vs. 0.5387), it does not always do so as the shot boundary detection technique works better on 21 out of the 62 days.

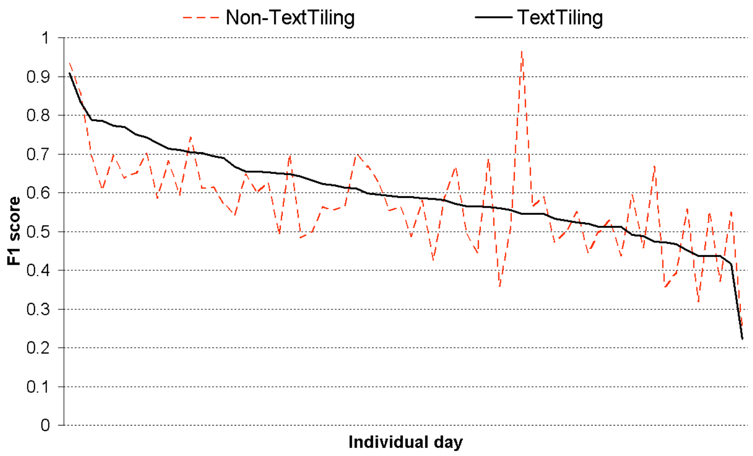


Figure 3.14: Performance benefits in using TextTiling on the image source of information.

3.3.3 Optimal Normalisation Technique

Given that there will be a list of difference scores computed for each individual source of information, a number of techniques were investigated so as to fuse the scores together. Given that no one source will always perform optimally, performance may be improved by combining multiple sources together. Before the results lists from the previous processing stages can be fused together, it is firstly necessary to normalise the lists to the same scale i.e. scores from the light sensor will be on a different scale to scores produced by the temperature sensor. We now investigate the performance of 4 standard data normalisation techniques described earlier: Standard, Min-Max, Sum, and Mean-Shift.

Mean-Shift performed poorly because many normalised scores were negative values (this problem was also experienced by Montage and Aslam in their experiments [103]).

There was little difference in the the Min-Max (0.5852) and Standard (0.5854) normalisation techniques. In fact the *Min-Max* and *Standard* techniques are basically identical as the datasets will generally have sensor or image similarity values of 0, hence the minimum score is shifted to 0 for the Min-Max technique (just like the Standard technique). Overall the *Sum* method performs best with an overall F-Measure score of 0.6271, however as illustrated in Figure 3.15 it isn't *always* better than the other techniques (on 21 out of the 62 days in the test set *Min-Max* normalisation outperformed *Sum*).

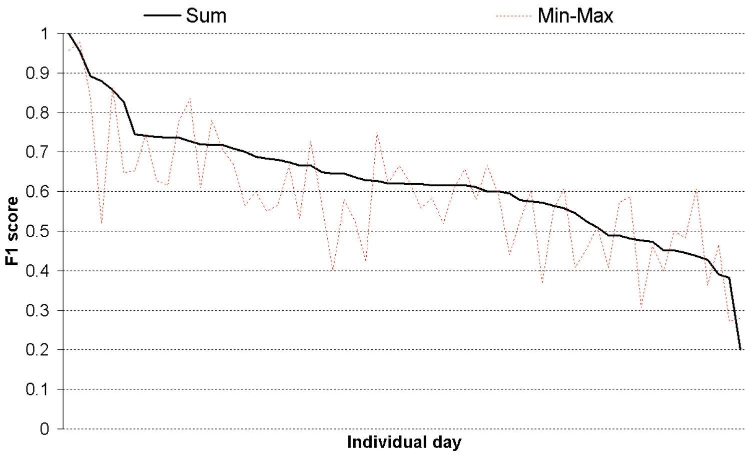


Figure 3.15: Breakdown of daily segmentation performances by normalisation approach.

3.3.4 Optimal Fusion Technique

Having determined that Sum normalisation is the best technique to put all the boundary likelihood scores from the various sources on to the same scale, we now investigate the optimal approach to fuse those scores together. As discussed earlier in this chapter we investigate 6 different fusion methods: CombANZ, CombMAX, CombMED, CombMIN, CombMNZ and CombSUM. Given that some sources may be more reliable than others, we also performed extensive training so as to assign weights, or confidence scores, to each individual source.

Figure 3.16 shows that individually all the sources performed comparatively, even if more emphasis/weighting/confidence is placed on each individual source when weighted with other sources. The MPEG-7 and accelerometer sources perform *slightly* better, but not to a great degree. After extensive training, where fusion weights were altered in steps of 0.05, a fusion of the image and accelerometer sources performed best with the weightings outlined in Table 3.4.

As the best trained system is only combining two sources of information the performances of the *CombMIN* and *CombMED* approaches will return the same score, as will the *CombMNZ* and *CombSUM* approaches. Therefore in this paragraph the following fusion method results are investigated: *CombANZ*, *CombMAX*, *CombMIN*, and *CombSUM*.

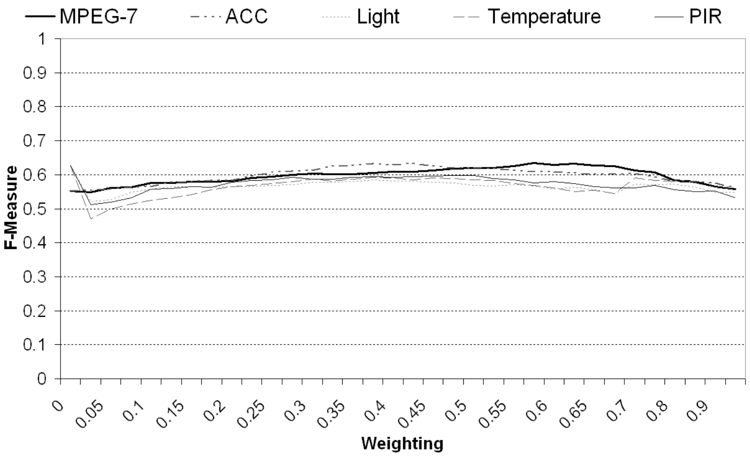


Figure 3.16: Effects on segmentation performance by increasing confidence (x-axis) on information sources.

Fusion Parameter	MPEG-7 + Sensors
Fusion Type	CombMIN
MPEG-7 Weight	0.65
Acc Combined Weight [38]	0.35
Acc X weight	0.0
Acc Y weight	0.0
Acc Z weight	0.0
Light weight	0.0
Temperature weight	0.0
Passive Infrared weight	0.0

Table 3.4: Summary of best trained fusion parameters for event segmentation.

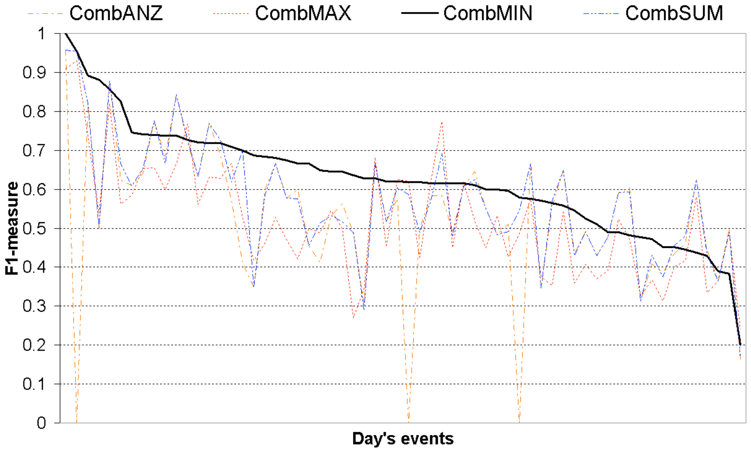


Figure 3.17: Breakdown of daily segmentation performances by fusion approach.

Overall, the optimal fusion approach is that of *CombMIN* (0.6271) which performs quite a bit better than the *CombANZ* (0.5430), *CombMAX* (0.5298), and *CombSUM* (0.5753) approaches. This approach is not *always* better though, as on inspection of Figure 3.17 it can be seen that the *CombMIN* approach is not the best performing approach 28.7% of the time (24 out of the 62 days of data in the test set). However, on those days that *CombMIN* isn't the best performing approach, it performs very close to the other approaches, only being on average 0.0216 worse than the best performing approach of the particular days in question.

3.3.5 Effects of Peak Scoring

After the scores from the various sources have been fused into a single list of event boundary likelihood scores, we now investigate if our proposed *peak scoring* technique can help improve event detection performance even further. Peak scoring exaggerates any peak values in the list of the scores and makes a graph of all the values appear more spiky (as illustrated back in Figure 3.8).

We found that on average it is better to use our proposed peak scoring method which boosts the F1-Measure figures for overall boundary identification from 0.5393 to 0.6271. Out of the 63 days of test data from the 5 users, on 60 of these days the peak scoring method resulted in better segmentation performance, as illustrated in Figure 3.18.

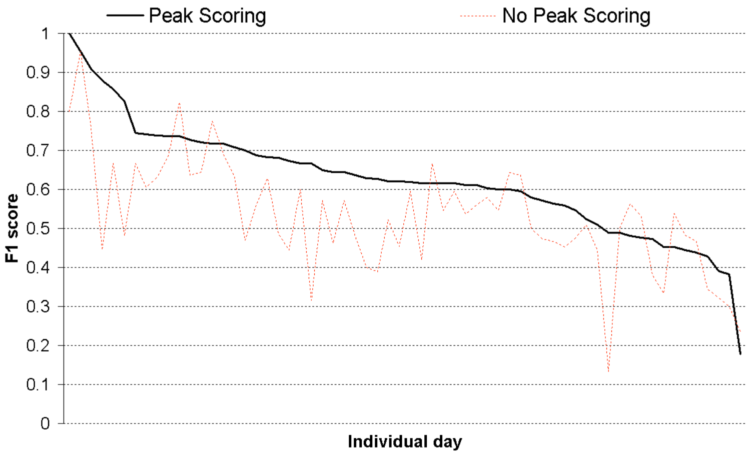


Figure 3.18: Effects on daily segmentation performances by using peak scoring or not.

It is in fact interesting to note that the peak scoring performs worse in terms of precision (Figure 3.19), but *always* better in terms of recall (Figure 3.20). This is due to the "spiky" nature of the output of the peak-scoring approach (Figure 3.8), in which peaks are exaggerated, thus proposing more event boundaries, which increases recall but can propose more false boundaries too, which has a detrimental effect on precision. However, in terms of the F-Measure, using peak scoring is much better.

3.3.6 Optimal Thresholding Technique

Given that the facets thus far produce a likelihood score that instances in the day are event boundaries, we still have to select a final set of instances as proposed event boundaries. As a result in this section we investigate three thresholding techniques to select either the top 20 likelihood values in a given day (RIAO approach), or to select a threshold value based on the likelihood scores and select only instances that score higher (Kapur and Mean methods).

While the thresholding approach used in an earlier publication of our work [50] performs best overall in terms of the number of true positives returned, and the Kapur method performs best in terms of recall, the performance of the mean thresholding approach performed best on the test dataset in terms of producing both high recall and precision values, as is evident in Table 3.5.

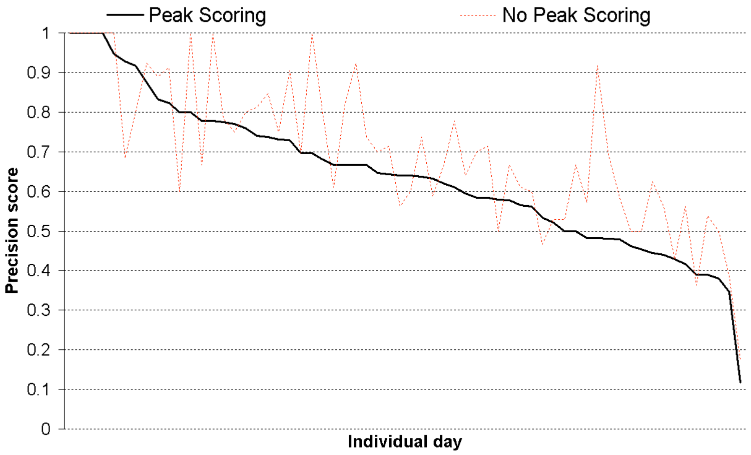


Figure 3.19: Effects on daily segmentation *precision* performances by using peak scoring or not.

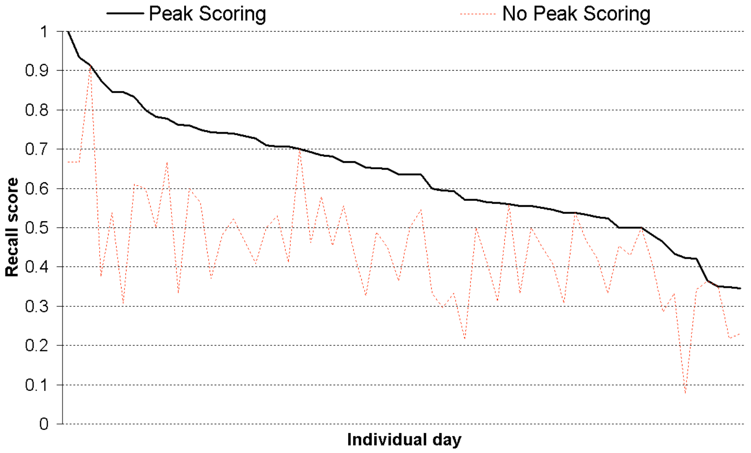


Figure 3.20: Effects on daily segmentation *recall* performances by using peak scoring or not.

Thresholding Method	Precision	Recall	F1-Measure
Mean (k = 3.4)	0.6294	0.6249	0.6271
Kapur (64 bins)	0.4891	0.7121	0.5799
RIAO (top 20)	0.6789	0.4642	0.5514

Table 3.5: Breakdown of daily segmentation performances by thresholding approach.

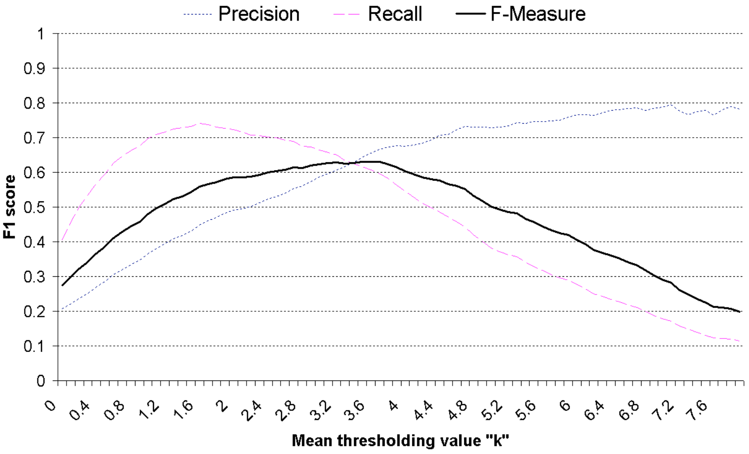


Figure 3.21: Performance effects of changing the mean thresholding parameter ”k”.

Observing the effects of changing the mean thresholding parameter on the test set *k* it can be seen in Figure 3.21 that as the threshold is increased precision is enhanced, but recall is much worse. In training we found a value of 3.4 for *k* was optimal.

3.3.7 Optimal Post-Processing Boundary Gap Method

Finally due to the nature of the data, where certain moments may have quite a bit of noisy data, some of the proposed events after thresholding may be too short in nature. As a result we carry out a post processing step to delete any events that are less than *x* minutes in length, and merge such events with the next event in question. This step is referred to as the post-processing boundary gap in this thesis.

As the magnitude of this parameter is increased, precision also increases, as the effects of over-segmentation (and thus false positives) is nullified (i.e. one false boundary triggered very soon after a true boundary will be ignored). However, this leads to a negative change in the level of recall too, as illustrated in Figure 3.22. Through experimentation we found that a gap of 3 minutes produced the best F1-Measure.

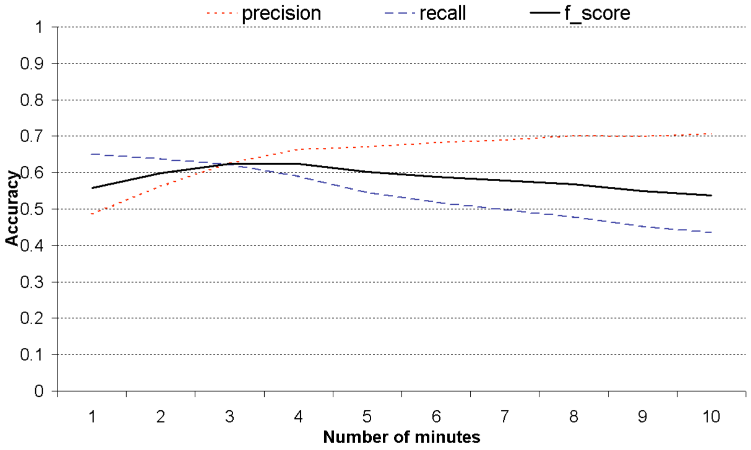


Figure 3.22: Performance effects of changing the post-processing boundary gap parameter.

Fusion Parameter	Sensors Only
Fusion Type	CombMIN
MPEG-7 Weight	**0.0**
Acc Combined Weight [38]	0.25
Acc X weight	0.35
Acc Y weight	0.0
Acc Z weight	0.0
Light weight	0.0
Temperature weight	0.0
Passive Infrared weight	0.4

Table 3.6: A summary of the best trained fusion weights using only the sensor sources of information.

3.3.8 Incorporating Only Sensor Sources

As stated earlier, it takes approximately 30 minutes to extract the MPEG-7 descriptor values from a busy day of 2,500 SenseCam images. Therefore it was decided to investigate the event segmentation performance of the system described using sensor sources alone. By using only sensor sources, event segmentation processing is just a few seconds of computation time. A fusion of the accelerometer and passive infrared sources (sensor sources only, hence extremely quick processing time), outlined in Table 3.6, provided the best trained parameters.

The overall event segmentation performance of this *Sensors Only* system was 0.6072 as

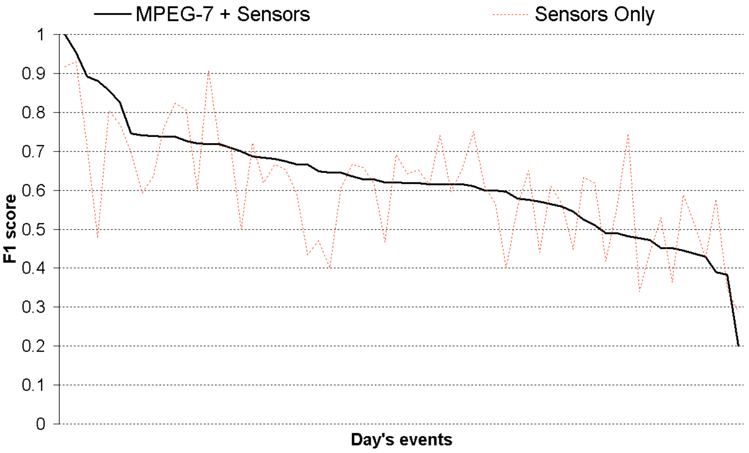


Figure 3.23: This graph illustrates the daily segmentation performances are comparable between using just the sensors and the best trained overall (MPEG-7 + Sensors) approach.

against a performance of 0.6271 by the best system which incorporates MPEG-7 sources. On inspection of Figure 3.23 the *Sensors Only* system (thin line) performs comparatively to the *MPEG-7 + Sensors* system, as it is better on 28 of the 62 days (the *MPEG-7 + Sensors* system is better on 34 days). It is interesting to observe that in terms of overall performance for each user, the *Sensors Only* system works better for 3 out of the 5 users (users 3, 4, and 5 in Figure 3.24). A possible explanation for this is that users 3, 4, and 5 generally wore their devices in a more office-bound environment. The users were working hard at desks most of the day, so any movement changes generally meant going for a meeting, going to lunch, going home, etc. Whereas users 1 and 2 had more active lifestyles in that they were less office bound, and also at weekends were quite active, therefore motion events detected by the sensors may not always have a semantic meaning to the user. However, as can be seen in Figure 3.24 the sensor based segmentation still works very good (relative to MPEG-7 + Sensors) for both users 1 and 2.

3.3.9 Comparison of System against State of the Art

Approaches of other research groups were implemented to allow the experiments in this thesis be compared against the current state of the art. This was done because it is imprac-

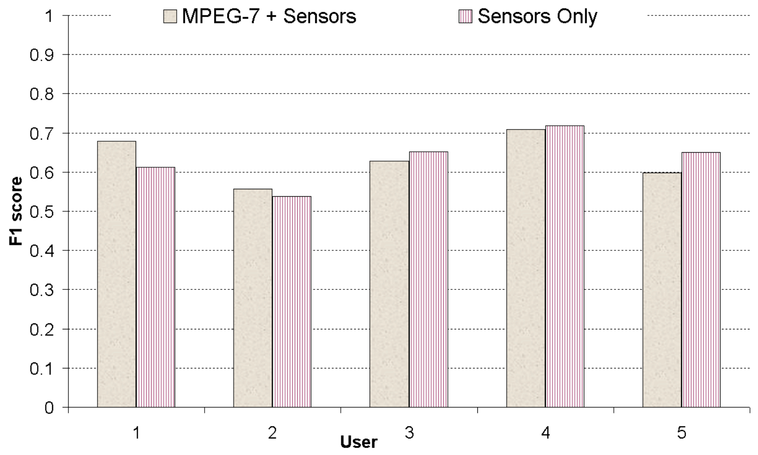


Figure 3.24: This graph illustrates the average performance on each user's data of the sensors only approach versus the best trained overall approach.

tical to apply our findings on the data of others and unreliable to compare performances across different data sources. The other approaches we implemented and compare against are:

- *Princeton* - Wang *et. al.* segment their lifelog of video into 5 minute clips of fixed length, however real events are not always 5 minutes in duration [155].

- *Yeung-Yeo* - Yeung & Yeo detail a time-constrained clustering technique for story boundary detection in the video domain [157]. A weakness of this approach is that the number of events/clusters must be fixed to a value *k*, and also only one data source is considered (in fairness Yeung & Yeo only had video footage available).

- *RIAO* - In earlier published work we focused on investigating what individual and combined data sources yield the richest source of information for activity segmentation [50]. However, the selection of a threshold for the number of events in a day was arbitrarily fixed at 20, and only the precision of the segmentation was computed as we had no detailed groundtruth of valid event boundaries from which to calculate recall.

- *Sensors Only* - The best performing system from earlier in this chapter which uses

Thresholding Method	Precision	Recall	F1-Measure
Sensors Only	0.6102	0.6043	0.6072
Yeung-Yeo	0.4318	0.5152	0.4698
RIAO	0.6488	0.3209	0.4294
Princeton	0.1910	0.9781	0.3196

Table 3.7: Overall comparison of event segmentation techniques against the prior state of the art.

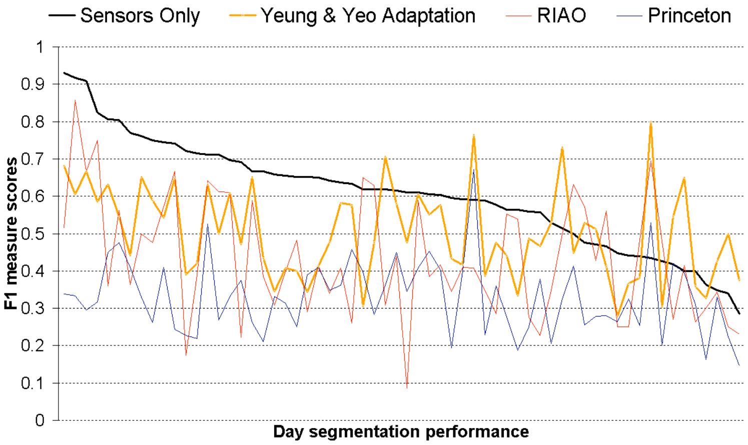


Figure 3.25: Breakdown of daily segmentation performances of our proposed approaches versus other approaches in literature.

only sensor sources.

Table 3.7 details the overall performance of each approach. It is observed that the approach proposed in this thesis performs much better in terms of the important F1-Measure than other approaches in the literature. Figure 3.25 shows the results on a day to day basis and it can be seen that the approaches from earlier in this chapter are consistently better than the previous known alternative. Our proposed approach performs better on 45 out of 62 days and 29.2% better overall than the next best system (Yeung & Yeo [157]).

3.4 Summary

In our hypothesis we have stated that we aim to *"... effectively segment a large collection of personal images into distinct events ... "* The work reported in this chapter was to investigate intelligent techniques to segment sequences of SenseCam images into distinct events or activities, which is important for later recall of memories and also helps to digest the large volume of personal images that can be collected. This challenge is somewhat analogous to the challenge of scene boundary detection in the video domain, but there are additional complexities inherent with segmenting sequences of lifelog images: 1) The fact that video is captured at approximately 25 frames per second, while lifelog images from the SenseCam are captured approximately just 3 times *per minute*, and 2) Lifelog images on the SenseCam are taken with a fisheye lens and are only 640 x 480 pixels in size, thus they are of low image quality. A number of questions are investigated in this chapter: attempting to determine the best vector distance technique to compare adjacent images against each other; whether to compare single adjacent images or blocks of adjacent images (TextTiling); the optimal normalisation and fusion techniques; whether to increase the likelihood of peak differences between adjacent (blocks of) images being event boundaries or not (peak scoring); the optimal normalisation technique to use; and determining the minimum amount of time that each event should last (post-processing boundary gap).

5 users each wore a SenseCam for one month, collecting a total of 271,163 images. The users then created a manual groundtruth of the images that were boundaries between different activities/events. In all 2,986 boundaries were identified, which is a test collection of considerable size.

The best trained system achieved a precision of 0.6294, a recall of 0.6249, and a F1-Measure of 0.6271. The best vector distance comparison method we found to use overall is the *Histogram Intersection* approach, even though it is not always the best method to use in each specific instance, it will perform best overall. The introduction of the TextTiling technique on the MPEG-7 source of information results in an 11.8% performance increase, while for the temperature sources it results in an improvement of 15.1%, and finally for the passive infrared source it results in massive improvements of 510.3%. However, for sources of information that can quickly change in values, it is optimal to only compare adjacent images e.g. the accelerometer source is 59.8% better when not using TextTiling, and the light source is 30.6% better when not using TextTiling. The optimal normalisation technique to use is the *Sum* normalisation approach, while the optimal fusion technique is that of *CombMIN*. After training it was decided to use just 2 sources of information: the MPEG-7 sources (early fused together, weighting of 0.65) and the accelerometer source (weighting of 0.35). The introduction of *Peak Scoring* results in a performance gain of 16.3%, and is thus recommended. The mean thresholding technique with a parameter of $k = 3.4$ is the best thresholding technique to use (8.1% better than the Kapur thresholding technique). Finally a *Post-Processing Boundary Gap* of 3 minutes was decided as the minimum length that each event/activity should be in duration.

Overall the incorporation of MPEG-7 + sensor sources offers an F1-Measure of 0.6271, however a system based on sensor sources only will perform *almost* as well (0.6072). Given the large reduction in processing time it is recommended to use this approach in systems that incorporate many users collecting significant amounts of data. This system has been proven to work better than other approaches in previous literature in this area of research; 29.2% better than Yeung & Yeo [157], 41.4% better than an earlier publication of our own [50], and 90% better than Wang *et. al.* [155].

3.5 Conclusions and Contributions

- We utilise an adaptation of Hearst's *TextTiling* technique [70] to compare "blocks" of adjacent images or sensor values. This technique performs optimally on sources of information where the values change slowly, i.e. the image, temperature, and passive infrared sensor sources. For those sources of information where the values can change very quickly it is better to compare adjacent readings i.e. the accelerometer and light level sensors.

- We investigated a number of vector distance techniques with no single technique performing significantly better than the rest. The simple *Manhattan* and *Euclidean* techniques perform almost as well as the best performing *Histogram Intersection* technique.

- We also investigated a number of normalisation and fusion techniques. We found the *Sum* normalisation technique worked best, while the *CombMin* fusion technique performed optimally.

- After the fusion stage it is highly recommended to use our *Peak Scoring* technique to exaggerate the large changes in activity.

- We discovered that a parametric based thresholding technique with a relatively high threshold (of 3.4 times the mean of all the values) has a large impact on final segmentation performance.

- No system should allow events to be less than 3 minutes in duration.

- Our most important discovery is that segmentation performance can be adequately achieved through the use of only onboard sensor values. This means that the costly process of extracting image features is not necessary and thus processing is performed significantly more quickly. It was found that a combination of accelerometer sensors along with the passive infrared sensor performed best among the onboard sensors. In fact, the nature of the passive infrared (PIR) sensor means that it is very sensitive to movement, and when the SenseCam is not stationary, the PIR sensor is in essence another motion sensor. To conclude we believe that only motion based sensors are needed to segment a lifelog of images into distinct events.

- We have published papers on event segmentation at the 2007 RIAO [50] and 2008 WIAMIS [48] conferences.

Chapter 4

IDENTIFYING SIMILAR EVENTS

> *In order for visual lifelogging technologies to effectively provide autobiographical memory retrieval cues it is necessary to automatically:*
>
> ...
>
> - *usefully provide the ability to link similar events together*
>
> ...

A user will, on average, collect 700,000 SenseCam images over an entire year which relates to an average of 7,000 events. For a user looking at a particular event (e.g. talking to a friend), he may desire to search for other similar events (owing to the fact that the human mind stores information in an associative manner) and review his past experiences by clicking and viewing on relevant items, which then inspires the user to look at other related events and browse through them. However, if a user is capturing approximately 8 thousand events per year, she will soon be overwhelmed by such a volume of information, therefore it will be useful to provide search/retrieval facilities. The experiments detailed in this chapter aim to identify the optimal techniques to usefully provide the ability to retrieve relevant events to any given SenseCam event, as illustrated in Figure 4.1.

Traditionally the lifelogging community has been focused on the miniaturisation of capture devices, and also on the problem of storing the very large amounts of data generated by these devices. In this chapter we firstly identify that the community are only now attempt-

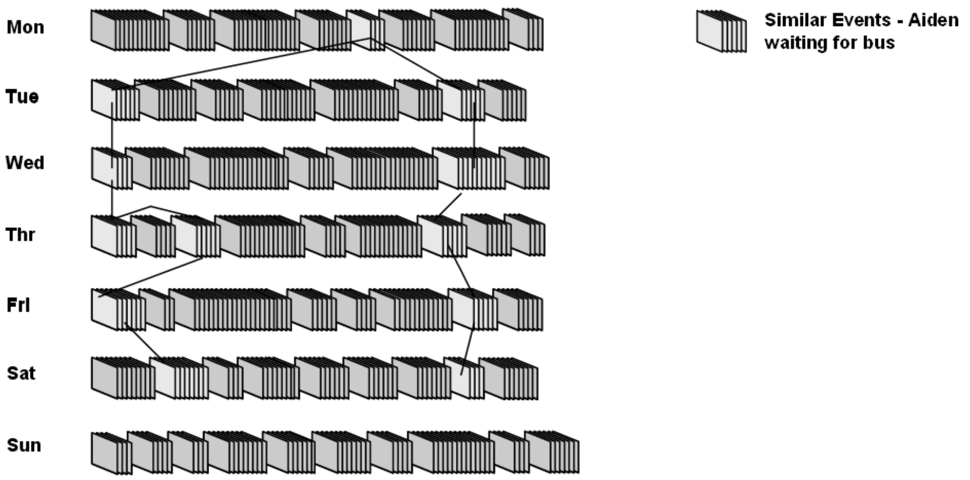


Figure 4.1: Retrieving other events that are similar to a given lifelog event.

ing to address the significant challenge of retrieving material from a lifelog that is relevant to a user's information need.

Retrieval in the domain of lifelogging has been investigated before, however experiments have been on very small datasets confined to the data of one user [155, 89, 146, 53]. Also, the retrieval systems have been based on content sources only [89, 53, 155], or context sources only [146, 31], and no lifelogging researchers have looked at combining content and contextual sources[1]. The combination of context and content to the area of retrieving lifelog data is the research contribution of this chapter.

In terms of managing huge volumes of SenseCam events, we allow users to filter and browse using the 3 axes of who, where, and when i.e. who was at the event, where was the event, and when was the event. To complement these 3 axes we can also filter lifelog content by using higher level semantic features which has shown great promise for filtering based on concepts other than named people [30]. However, whether one's collection of lifelog events is filtered or not, there is a need to either browse all events or to do searching and event matching. In the case of the latter it is necessary to determine techniques to identify the optimal content-based system to retrieve other relevant events to any given SenseCam

[1] In using the words "content" and "context" we are somewhat influenced by our background in video retrieval, whereby we mean that "content" is features that can be derived from image pixels or audio recordings, and "context" is any other sources that can be gathered, e.g. SenseCam sensor values.

event. This case is illustrated in Figure 4.1 where for a search scenario on the 5^{th} event of the first day, all similar events from other days are linked to it.

To retrieve events similar to a given event in a lifelog it is necessary to firstly determine *how to represent* SenseCam events, and then *how to compare* those event representations against each other. We compare our proposed approach to a selection of other lifelogging retrieval methods. Two datasets are used, one of 273,744 SenseCam images with a groundtruth which also allows the tuning of retrieval parameters, and one of 1,864,149 SenseCam images where users judged retrieved events to a given query event for each approach. We discuss how we address the challenges of retrieval in lifelogs throughout the remainder of this chapter.

4.1 Approaches to Finding Similar Events

We will now discuss possible approaches towards retrieval of SenseCam events. The aim is to find similar events to a given event, as displayed in Figure 4.1. However, at the most basic level an event will consist of many images. Therefore the first question is how to represent an event? The next question can then be asked; how to compare those event representations against each other to find similar events?

Investigations could be carried out looking at using different vector distance, normalisation, fusion, and individual source weightings for representation and search variables, however our work should use the same value for both (representation and search) variables. If Manhattan distance is the best distance metric to compare individual SenseCam images against each other (to represent the events), then it is probably also the best at comparing whole event representative values against each other (to search for similar events).

We now detail our approaches towards representing SenseCam events, and then how we retrieve similar events to a given event by using those event representation values.

4.1.1 Event Representative Approach

On average each SenseCam event consists of almost 100 images. Each of those images is represented by a vector value, however it is desirable to obtain a single vector that is representative of the values of all 100 vectors. Therefore how should each event be represented? There are a number of approaches that we investigated, including:

1. **Middle Image** - Use the middle image from the event

2. **Event Average** - Get the average value of all the features across all the images in the event

3. **Exhaustive Within Event Keyframe** Select the image within the event that is closest to all the images in the event. To select an event representative image Cooper & Foote select the image that is closest to all other images in the event which requires ($n * n$ comparisons, where n is the number of images in a given event) [41].

4. **Within Event Keyframe** - Select the image within the event that is closest to the *average* value of all images in the event.

5. **Exhaustive Cross Event Keyframe** Select the image within the event that is closest to all the other images in the event, but most different to all the other images in all the other events of the same day. Cooper & Foote discuss this method to select an event representative image [41], resulting in a processing load of $n * m$ where m is the number of images in a day.

6. **Cross Event Keyframe** - Select the image within the event that is closest to the *average* value of all the images in this event, but most different to the *average* value of all the images in the other events of that same day. This reduces processing from $n * m$ to $n * e$, where m is the number of images in a day, and e is the number of events in a day, with [$e << m$, typically $m = 90 * e$].

7. **Quality Keyframe** - Select the image within the event that has the highest image “quality” score. This image may be a good representation of the semantic meaning

of the event. We discuss the notion of image "quality" in much detail in the next chapter.

8. **Middle N** - Select the average value of all the features across the middle *n* images in the event. Given that sequences of SenseCam images can be segmented into distinct events or activities quickly using sensor values, we investigate extracting the MPEG-7 image descriptor values only from a selection of images in the middle of each event. The premise of this is that images in the middle of an event are more likely to be representative of the semantic meaning of that event. Also by taking the average of a number of images, the effect of single poor quality or outlier images is less damaging on retrieval performance.

If using the *Cross Event* or *Exhaustive Cross Event Keyframe* representation technique there is a trade off as to how much importance to place on the image that is closest to all the images (or average) within the same event, as opposed to the image that is most different to all the images (or average) from all the other events. This trade-off will be empirically determined.

4.1.2 Weighting of Images towards Middle of Event

When comparing images against each other to determine the event representative image, or when selecting the average features of all the images to represent the event, the benefits of weighting more strongly those images towards the middle of the event will be investigated. The rationale behind this is that those images at the start and end of the event may in fact belong to the previous/next events (as event segmentation may not always be perfect) or may belong to a transition from/to the previous/next events. Therefore it is investigated whether linearly weighting images towards the middle of an event is worthwhile.

4.1.3 Vector Distance Approach

When using the “within” or “cross” event selection images from the representation-keyframe approach it must be determined how to compare images against each other, and also how to compare event representative images/vectors against each other. The MPEG-7 features of each image (or event representation) are represented as a vector. Therefore the optimal vector distance comparison method is investigated for these tasks. The following methods (which are described in Chapter 3) are investigated: *Bray-Curtis*, *Canberra*, *Euclidean*, *Histogram Intersection*, *Jeffrey Modification of Kullback-Leiber*, *Kullback-Leiber*, *Manhattan*, *Square Chi Squared*, *Squared Chord*, and X^2 *Statistics*.

4.1.4 Fusion of Data Sources

As there are a number of information sources available (MPEG-7 from images , accelerometer, light, passive infrared, and ambient temperature values) it is necessary to investigate the optimal normalisation and fusion techniques. Normalisation is necessary as the scores from the particular data source are on different scales, and it is therefore required to normalise these sets of scores to be on the same scale. By doing this, it is then possible to combine various sources of information together. The following normalisation techniques, explained in the segmentation chapter, are investigated: *Mean-Shift*, *Min-Max*, *Sum* and *Standard*. After normalisation, it is then possible to combine the results from the various data sources which may offer more reliable retrieval results. A number of possible fusion methods, also detailed in the previous chapter, are investigated: *CombANZ*, *CombMAX*, *CombMED*, *CombMIN*, *CombMNZ*, *CombSUM*.

Given that certain sources of information may be more reliable in terms of comparing any 2 given lifelog events to determine how similar they are, we empirically train parameters to investigate which sources of information should be used, and how much confidence should be attached to each source e.g. it may be desirable that more emphasis should be placed on the MPEG-7 source when comparing events to the reference event, i.e. we want to put more emphasis on events that are visually similar to the query event, as opposed to say those that are of a similar ambient temperature.

As explained in Chapter 3, there are a multitude of available MPEG-7 descriptors, and we have selected 4 that are well suited towards SenseCam images. To recap, the MPEG-7 descriptors we use are colour layout, colour structure, scalable colour, and edge histogram [20]. It will be investigated which of those sources of information are most useful (when early fused) in comparing event representations against each other.

4.1.5 Summary of Approaches to be investigated

In this section we have detailed that to retrieve similar events to a given lifelog event, it must firstly be determined how all events should be represented. On average each event consists of almost 100 images, and we will investigate a number of techniques to select a single representative image, or an average value of all (or a selection of) the images. We also investigate whether it will be useful to more strongly weight those images towards the middle of a given SenseCam event, on the premise that they are more likely to represent the semantic meaning of events.

It is only then, after deciding on how to represent all the events in an individual's lifelog collection, that it is possible to compare all events against each other to determine how similar they are. Given that MPEG-7 image values are represented by vector values, we will investigate the optimal vector distance comparison technique to use in the retrieval of lifelog events. Also given that there are a number of different sources of information available to represent all the events, it is necessary to investigate combining these sources together. This involves investigating various normalisation and fusion methods, as well as investigating how much confidence to place on the various sources of information.

We now move on to the next sections which describe experiments we set up to investigate the effectiveness of our proposed approaches, and how the various facets of the system perform on real user queries.

4.2 Experimental Setup

In experiments to investigate the effectiveness of our retrieval approaches, we asked 5 users (information retrieval specialists aged 25-35) to collect SenseCam data. We collected two datasets, one of 273,744 images, and another of 1,864,149 images. The first and smaller dataset was used to construct an extensive groundtruth of relevance judgements. To have a sufficient number of relevant events to train parameters on, it was decided to go for more general queries in this dataset e.g. driving, at work on PC, eating, etc. The purpose of the second and larger dataset was to investigate how the performance of the best systems, trained on the smaller dataset, translates to an extensive lifelog of images. Also the queries used on the larger dataset were much more specific e.g. "what other times was I talking to John?", "what other times was I on an aeroplane?", etc. We will now describe the experimental setup of firstly the smaller dataset with the relevance judgements, and secondly the very large dataset with the very specific user-generated queries.

4.2.1 Dataset with Relevance Judgements

In experiments to investigate the effectiveness of our retrieval approaches, we asked 5 different users to collect SenseCam data over a period of 30 days. A total of 273,744 SenseCam images were used in this experiment (all images of less than 4KB in size were ignored, as these are images of total darkness, i.e. when SenseCam is behind a coat, etc.). These images were segmented into events using the optimal segmentation approach identified in Chapter 3 on event segmentation. Table 4.1 summarises some of the main statistics broken down by user.

Each and every image has sensor values associated with it, and also MPEG-7 descriptor values were extracted. It takes approximately 30 minutes to process a larger than average day of 2,500 images (on a 2.4GHz Pentium 4 machine with 512MB RAM). Therefore to process all the images it took approximately 75 hours (150 days data, and 30 minutes to process each day).

User	Total Num Images	Num Events	Avg Events Per Day	Avg Images Per Day	Images Per Event
1	79,595	1,071	30	2,274	74
2	76,023	892	19	1,584	85
3	42,700	409	19	2,033	104
4	40,715	492	20	1,629	83
5	34,711	422	18	1,509	82

Table 4.1: Statistics on data collected for the “general” dataset.

User	Num judgements to make
1	6,728
2	6,526
3	4,327
4	4,859
5	4,577

Table 4.2: The number of judgements to make going if we were to use the traditional pooling approach.

4.2.1.1 Constructing a Groundtruth

10 diverse query events were selected for each user, thus giving a total of 50 queries/topics. The users were then asked to judge a large number of potentially relevant events against each query event to build up a groundtruth of data. This was done in a TRECVid style pooling approach [154]. In calculating the events to be judged 43 possible system variations were output in which the top 100 results from each run (i.e. a pooling depth of 100) were stored. If the users were asked to evaluate all the (pooled top 100) returned results, for each of the 10 topics on the 43 system variations, it would result in 43,000 judgements to be made by each user. However, by removing duplicates this still requires an average of 5,404 judgements from each user (Table 4.2):

This would result in a large annotation burden on individual users; therefore it was investigated if any possibility existed to have the users judge on fewer results but to still keep *almost* all the relevant results. The results to be judged are given to the user if they appear in any of the pooled results from the 43 system runs, then it was investigated to see how many would be retrieved if users judged only those results that were returned by a number of the pooled systems (see Table 4.3). This was investigated on a sample query

Num systems proposing event (in top 100)	Unique events returned	Num correct events	Precision	Recall	F Score
No Pooling	891	106	0.119	1.000	0.213
1	479	93	0.194	0.877	0.325
2	363	85	0.234	0.802	0.373
3	**296**	**80**	**0.270**	**0.755**	**0.411**
4	266	75	0.282	0.708	0.418
5	239	63	0.264	0.594	0.380
6	221	57	0.258	0.538	0.363

Table 4.3: The effects of specifying the number of systems that must propose an event as relevant in the pooling process.

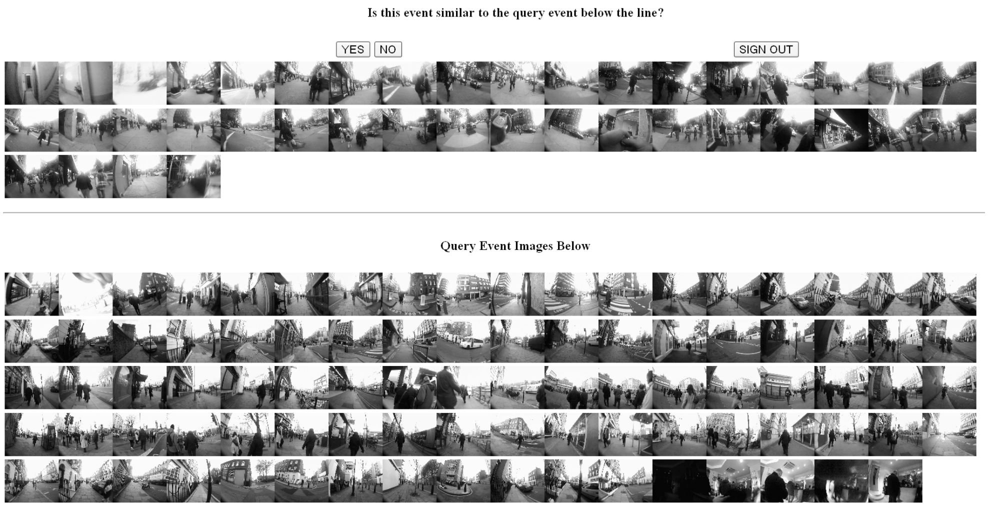
User	Num judgements to make	Original Number of judgements from Table 4.2
1	3,192	6,728
2	3,913	6,526
3	3,309	4,327
4	3,568	4,859
5	3,655	4,577
Total	17,637	27,017

Table 4.4: The final number of judgements that were requested from our users on the "general dataset" pooled list.

event from user 2 in Table 4.1 in which he was working on his computer.

It was eventually decided to proceed with the bolded row in Table 4.3, i.e. if a user is to judge any result, it will have to be proposed by at least 3 of the 43 different pooled systems. This places the least amount of judgment effort on the users, while still having a high recall figure from which to construct an extensive groundtruth to carry out the experiments that will be detailed in due course. This reduces the average amount of judgements expected from each user from 5,404 to 3,528 (see Table 4.4).

Figure 4.2 displays a screenshot of the application we developed where users judged the similarity of the pooled list of potentially relevant events against a given query/topic event. This application allowed a rapid form of annotation. The top section of the screen is updated with 40 sample images from a new event after the user clicks on the "yes" ("y") or "no" ("n") mouse/keyboard buttons to judge the relevance of the current event. After all images have been annotated for a given query event, the entire page is refreshed with a new



Is this event similar to the query event below the line?

Query Event Images Below

Figure 4.2: A screenshot of the application on which our users judged the similarity of potentially relevant events to reference events.

query event appearing below the line (100 sample images are displayed for query events since they are not constantly updated). The user is notified that they are annotating results for a new query event.

As an interesting point of reference there were "only" 7,393 results judged in TRECVid 2001 (which was a workshop of TREC that year), although these were judged by 2 different assessors [142]. This doesn't need to be done with SenseCam data as the only true judge can be the wearer of the individual device itself, due to the highly personal nature of SenseCam images. In TRECVid 2002 just 1 assessor judged the shots (as the '01 agreement between assessors was very good, 84.6%) giving a groundtruth of 23,681 judgments [140]. In TRECVid 2003, 39,077 judgments were made in the search task [139]. The experiments in this thesis are carried out on a groundtruth of 17,637 judgments from 5 different users across 5 different datasets. Compared to early TRECVid systems, the scale of our groundtruth is not insignificant, thus giving the retrieval results in this chapter a greater degree of gravity.

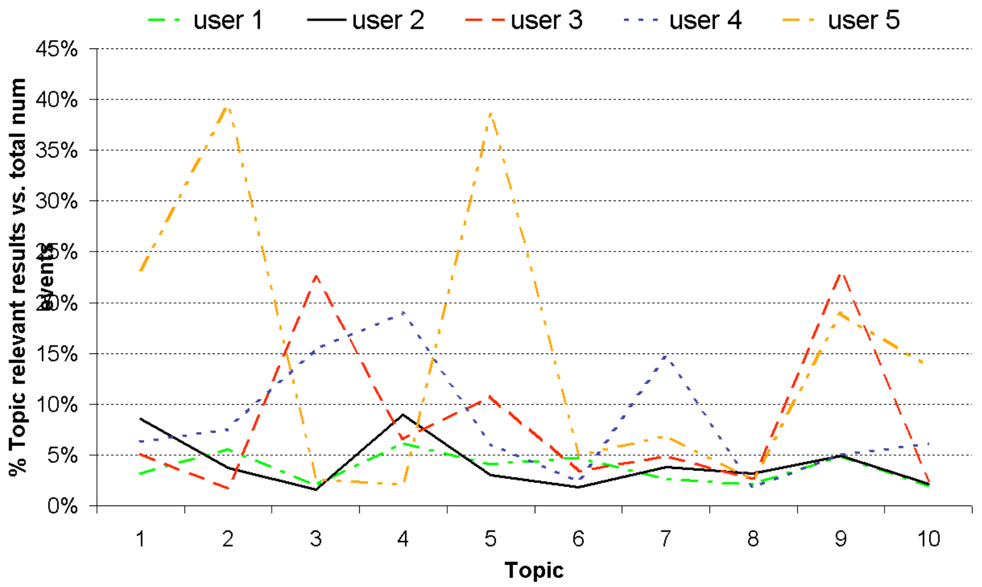


Figure 4.3: This unsorted graph illustrates the concentration of relevant results that each topic has associated with it (unsorted).

4.2.1.2 Dividing Data into Training & Test Sets

After the groundtruthing stage the 50 queries (50×5 users) were then divided into a training and test set. Taking the queries in the order that they were judged we get the distribution of relevant events as illustrated in Figure 4.3: The y-axis percentage values indicate the number of returned relevant result events for this query against the total number of events in the whole database for the given user.

However to select training and testing sets it wouldn't be correct to just take the first 6 or 7 queries as the training set and the last 3 or 4 as the test set, as there may be an uneven/unfair balance in the distribution in the number of relevant results contained in each. Therefore it was decided, for each user, to sort the queries by those that returned the most relevant results, as illustrated in Figure 4.4:

To split into training and test sets it was decided to follow the sampling method outlined in Table 4.5). This will therefore result in each user providing 6 queries for training, and 4 queries for testing. In terms of overall system judgement this means that there will be 30 queries (6 x 5 queries) for training and 20 queries for testing purposes. Another possible

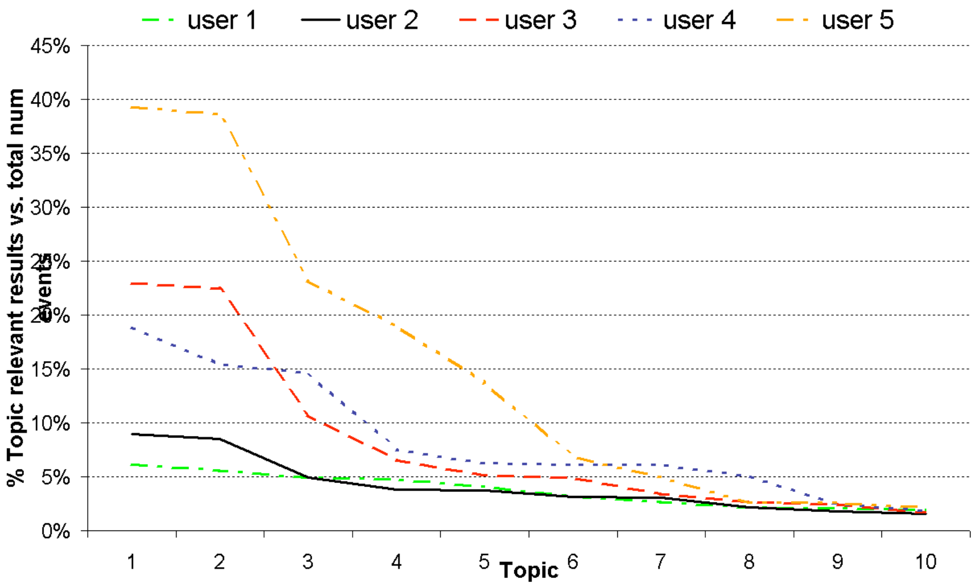


Figure 4.4: This unsorted graph illustrates the concentration of relevant results that each topic has associated with it (unsorted).

method to divide the data into training and test sets is to use n-fold cross validation which is useful when data is sparse [106]. However, in tests carried out in comparing both methods, with training time for the n-fold cross validation exceeding 30 hours, there was found to be only a 1.8% difference in both approaches with a very strong correlation of 0.99. For the sake of simplicity we use the aforementioned method of 30 fixed queries for training and 20 fixed queries for testing, rather than the more complicated method of n-fold cross validation.

4.2.2 Larger Dataset with Specific User-Generated Queries

A disadvantage of the dataset used in section 4.2.1 is that while it was necessary to select very general queries to produce a sufficient number of relevant events on which to tune retrieval parameters, these queries are not representative of *all* possible user query classes. Therefore we decided to create a second dataset on which users were asked to construct real world queries with very specific information needs. This dataset is also much more extensive in terms of size, thus adding to the challenge of producing good retrieval results,

Query Rank (Figure 4.4)	Training/Test Set?
1	Test
2	Training
3	Training
4	Test
5	Training
6	Training
7	Test
8	Training
9	Training
10	Test

Table 4.5: Our selection method for dividing queries into training and test sets.

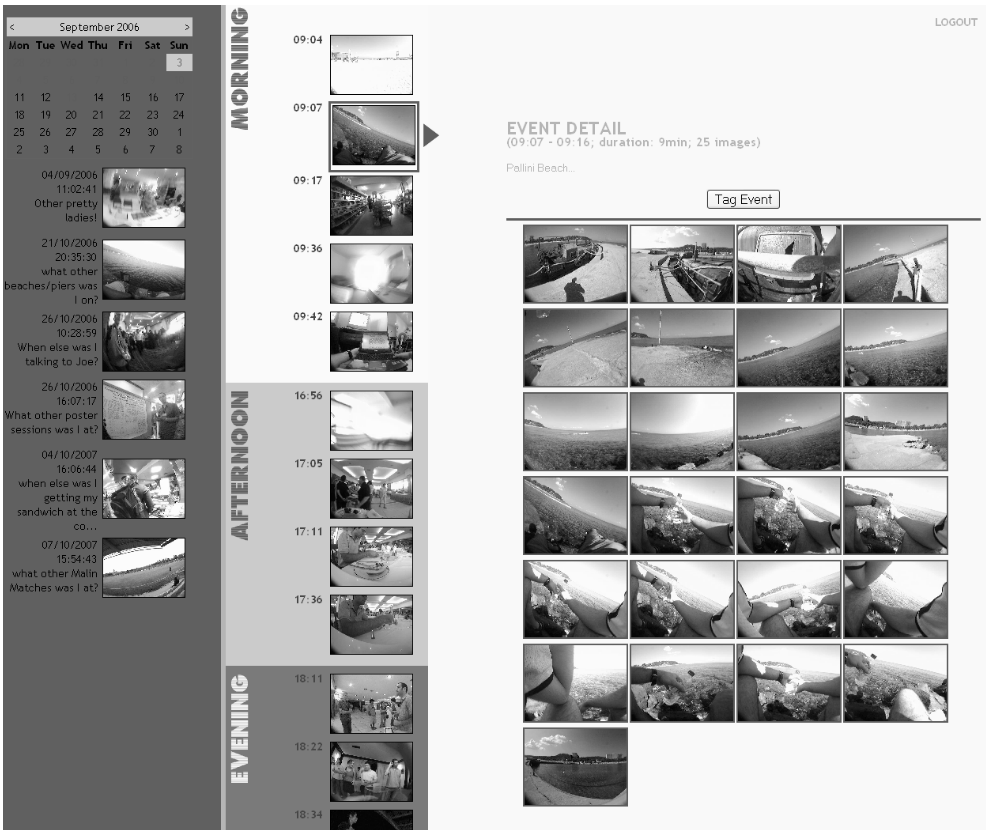
User	Total Num Images	Num Events	Num Days	Avg Events Per Day	Avg Images Per Day	Images Per Event	Num Queries
1	1,686,424	19,995	614	33	2,747	84	5
2	92,837	1,182	60	20	1,547	79	6
3	44,173	443	21	21	2,103	100	5
4	40,715	505	23	22	1,770	81	7

Table 4.6: Statistics on data gathered for the "specific" dataset.

but which also provides a more realistic evaluation of how such systems may perform in the real world.

In experiments to investigate the effectiveness of our retrieval approaches for real user-generated queries on extensive datasets, we asked 4 users to collect SenseCam data over a long period of time. A total of 1,864,149 SenseCam images were used in this experiment. These images were segmented into 22,125 events using the optimal segmentation approach identified in Chapter 3. Table 4.6 summarises some of the main statistics broken down by user. As opposed to the "general" dataset where 4 MPEG-7 image descriptors were extracted for each image, in this "specific" dataset two MPEG-7 features (scalable colour and edge histogram) were extracted for all of the images, taking an average of 13 minutes to process a typical day of 2,500 images, thus taking a total of 156 hours.

Users were asked to select a number of query events that they would like to see a list of potentially similar events to. They were presented with an event based browser (as



04/09/2006 11:02:41 Other pretty ladies!

21/10/2006 20:35:30 what other beaches/piers was I on?

26/10/2006 10:28:59 When else was I talking to Joe?

26/10/2006 16:07:17 What other poster sessions was I at?

04/10/2007 16:06:44 when else was I getting my sandwich at the co...

07/10/2007 15:54:43 what other Malin Matches was I at?

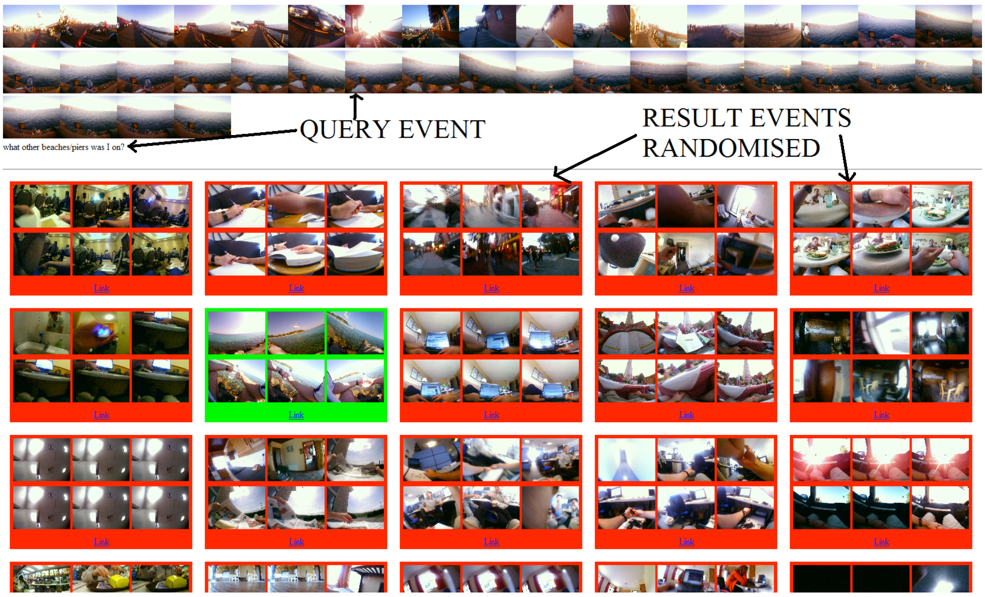
EVENT DETAIL
(09:07 - 09:16; duration: 9min; 25 images)

Pallini Beach...

Figure 4.5: A screenshot of the browser we used to mark events for annotation in the "specific dataset".

illustrated in Figure 4.5) to sift through their SenseCam images. The calendar allows the investigator to browse to a day of interest. The vertical column of images in the centre then displays each event for a selected day, and once clicked all images from the event are shown on the right of the screen. The users can select events to be query events by clicking on the "tag event" button. A list of tagged events is displayed under the calendar on the left hand side of the screen. In total, 23 events were selected to be used as queries (see "Num Queries" column in Table 4.6).

The retrieval of potentially relevant events was then processed for each query event in our dataset. Users were presented with a screen full of candidate events (20 from each approach) in which they had the option to select other events relevant to their lifelog query event, which were in turn highlighted with a green background (see Figure 4.6). The 6



what other beaches/piers was I on?

Figure 4.6: Screenshot of the application used to relevance judgements on specific user-generated queries.

keyframe images chosen for each candidate event were spread evenly e.g. first is image 16% through event, second 33%, third 50%, etc. Below each candidate image there is a "Link" button which opens all the images of that event in a new browser window. 1,736 unique candidate events were retrieved for presentation to users for judgement.

4.3 Evaluation of Event Retrieval Approaches

In this chapter we introduce the importance of retrieving relevant lifelog events to a given event. We have proposed a number of approaches to address this challenge. In this section we now detail our findings on the effectiveness of our approaches towards retrieving relevant events. The dataset of "general" queries with associated relevance judgements allowed us to tune various system parameters, and also to investigate the usefulness of various facets of the retrieval system.

In the training phase, an appropriate evaluation metric must be chosen on which to automatically tune various parameters. For the training phase of the segmentation experiments (see Chapter 3) the F-Measure was identified as the most important metric. However, because we are dealing with the challenge of retrieval in this section it is more important to have many relevant results returned with a high ranking, therefore measures such as precision @5 or precision @10 become important. Precision @N are measures relevant to this challenge, however they don't take into account the number of relevant documents in total, which is important for users who like to browse through many results. Standard information retrieval precision and recall measures are not suitable either, as they do not consider the ranking positions of the relevant results. In addition this dataset has an average of 44 relevant results per query, and if a high number of results are returned for every query (say 200), then it is quite likely that many systems will return these 44 relevant results, and therefore the overall recall and precision scores will always be the same across runs. The F-Measure is derived from precision and recall, so therefore it suffers from the same problem of not considering the ranked positions of relevant documents/events.

Eventually it was decided, in the training phase at least, to train based on the system that gives the best Mean Average Precision (MAP) score of the top 200 documents. The main argument for using MAP is that it takes into consideration the ranking of relevant results, with higher ranked relevant results being scored more highly than the same number of lowly ranked relevant results. However, it also takes into account the total number of relevant documents that can be obtained for a given query too. This is a desirable trait, and this is reflected by the importance placed by the TRECVid community on the MAP evaluation metric. MAP is calculated as the mean of the average precision scores for all the queries. Average precision for the results returned for a given query is calculated as follows in Equation 4.1:

$$AveP = \frac{\sum_{r=1}^{N}(P(r) - rel(r))}{R_N} \tag{4.1}$$

where R_N = total number of relevant documents

The best trained parameters (see Table 4.7) performed with a mean average precision of 0.3608 on the 20 test queries over all 5 users. A summary of the results is available in Table 4.8. The best trained parameters perform better on the test set than in the training

Research Variable	**Parameter Value**
Keyframe selection	event average
Vector distance	Canberra
Weight middle images	No
Normalisation	Min Max
Fusion	Comb MAX
Within/cross event weight	N/A
MPEG7 features	colour structure, scalable colour, edge histogram
MPEG7 weight	0.293
Accelerometer weight	0.057
Light weight	0.203
PIR weight	0.197
Temperature weight	0.25

Table 4.7: A summary of the best trained parameters in the “general dataset”.

Mean Average Precision	0.3608
precision	0.1748
recall	0.7021
f-Measure	0.2798
p @ 5	0.6000
p @ 10	0.5000
p @ 20	0.4375
p @ 30	0.3983
p @ 40	0.3725
p @ 50	0.3480

Table 4.8: A summary of how well the system with the best trained parameters performed on the test set.

set (MAP of 0.3381), an explanation for this is the fact that, on average, the test set topics contain a slightly greater concentration of relevant results due to the method of segmenting the queries into training and test sets (see Table 4.5).

We will now investigate the effects on the final retrieval performance of various facets explained earlier in this chapter.

4.3.1 Optimal Event Representative Approach

As lifelog events consist of almost 100 images on average, an appropriate technique must be chosen to represent each event using a single set of values. Then using these event representative values we are able to compare events with each other to determine how similar they

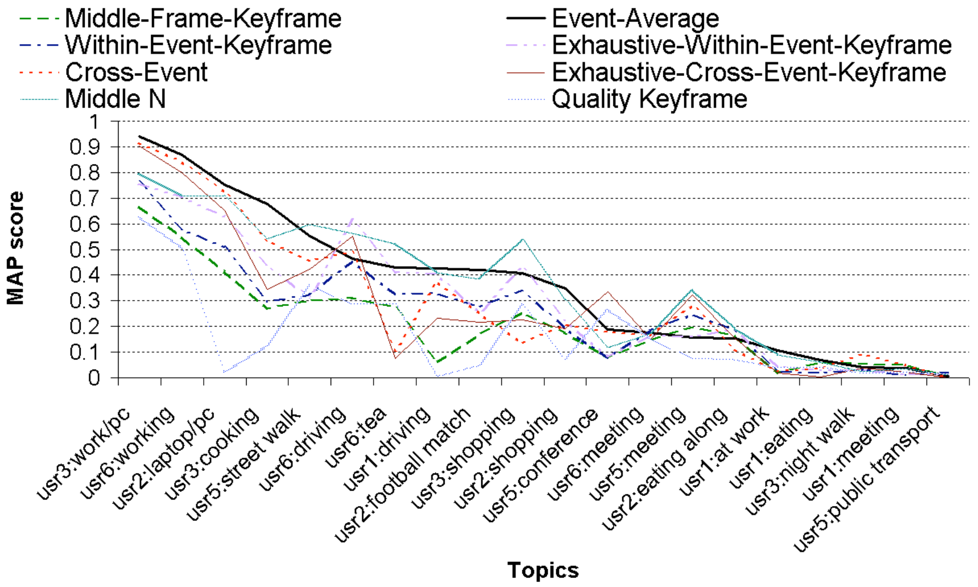


Figure 4.7: An examination on the retrieval performance affects by changing the event representation method on the test set. Topics are sorted based on their performance using the "Event-Average" approach.

are. Earlier in this chapter we described various techniques to represent an event, ranging from choosing an individual image to selecting the average value from a number of images.

The best trained parameter regarding the selection of a keyframe to represent an event was the approach of calculating the average value of all the images within the particular event for each source i.e. getting average of all the bin values of MPEG7 features, the accelerometer source, the passive infrared source, the temperature source and the light source. It can be seen that across many of the topics the *"Event Average"* approach works best (in 9 out of 20 topics as illustrated by the thick black continuous line in Figure 4.7). The *"Cross Event"* approach works quite well too, but it is much more computationally expensive than the *"Event Average"* approach (115 seconds vs. 42 seconds total processing time)[2].

Investigating if these results hold true across all users, Figure 4.8 displays that the *"Event Average"* approach performs best across all users too.

[2] Total processing time refers to the time it takes to represent all events in the database and then to investigate the similarity of each and every event against the query event, and finally to evaluate the performance of the ranked list against the groundtruth). It should be noted that this is performed over a very large database, and the retrieval part of this processing consumes a very small amount of time. This processing was carried out on

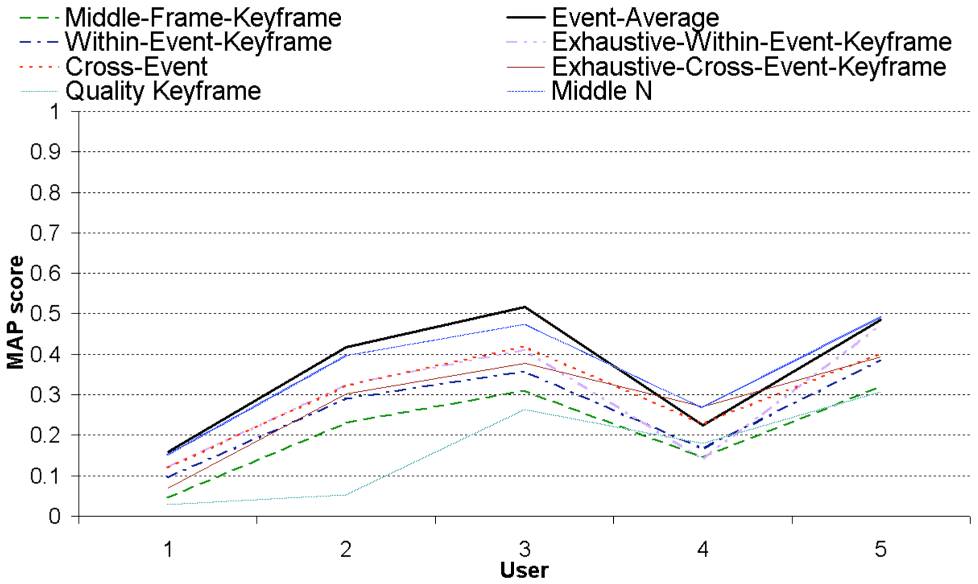


Figure 4.8: An examination on the retrieval performance affects on each user's dataset by changing the event representation method on the test set.

On the "specific" topics, from the larger dataset of 1.86 million images, the difference between the *Event Average* and *Middle Image* approaches is not as pronounced (0.217 P@10 for *event average* vs. 0.170 for *middle image keyframe*). However, still on 14 of the 23 topics the *event average* approach performs best, and overall we believe this is the superior approach.

4.3.2 Benefits of Weighting Images towards Middle of Event

Earlier in this chapter we rationalised that images in the middle segment of an event are more likely to be representative of the semantic meaning of that event, than those images towards the tail end of events. Therefore we investigated to see if linearly weighting images in the middle of events more strongly leads to increased retrieval performance.

On the test set it is only slightly better to weight images towards the middle of an event (MAP of 0.3666 vs. 0.3608). In considering Figure 4.9, it is apparent that over a range of topics the performance difference is almost negligible. Indeed the best trained parameter in

a 2.4GHz Pentium Dual Core machine with 2GB RAM.

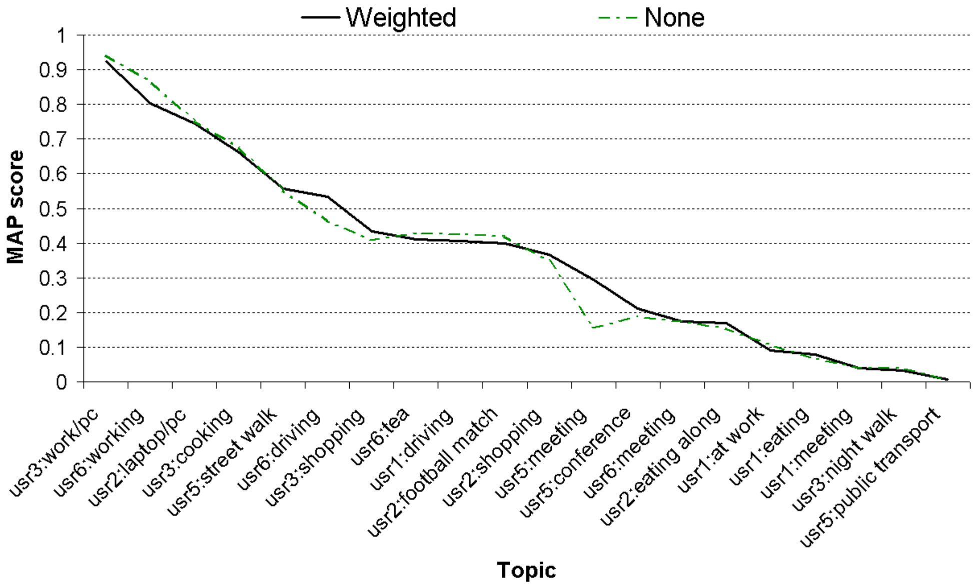


Figure 4.9: An examination on the retrieval performance affects on the ”general queries” test set by changing whether images should be more strongly weighted towards the middle of an event or not.

the training phase is not to weight images towards the middle (MAP score on training set of 0.3374 vs. 0.3240).

4.3.3 Vector Distance Evaluation

Given that we represent an event by the average of all its features, it must then be determined how to compare event representations against each other to see how similar they are. The sensor sources of information are represented by single scalar values, and this means that a simple difference value is appropriate. However, as the image source of information is represented by a vector value, there are a number of possible techniques that can be used to determine how similar two event representative vectors are. We now discuss our findings on the optimal such technique for comparing events against each other, to see how similar they are.

The best trained parameters indicate that the colour structure, scalable colour, and edge histogram MPEG-7 descriptors should be included, while the colour layout descriptor

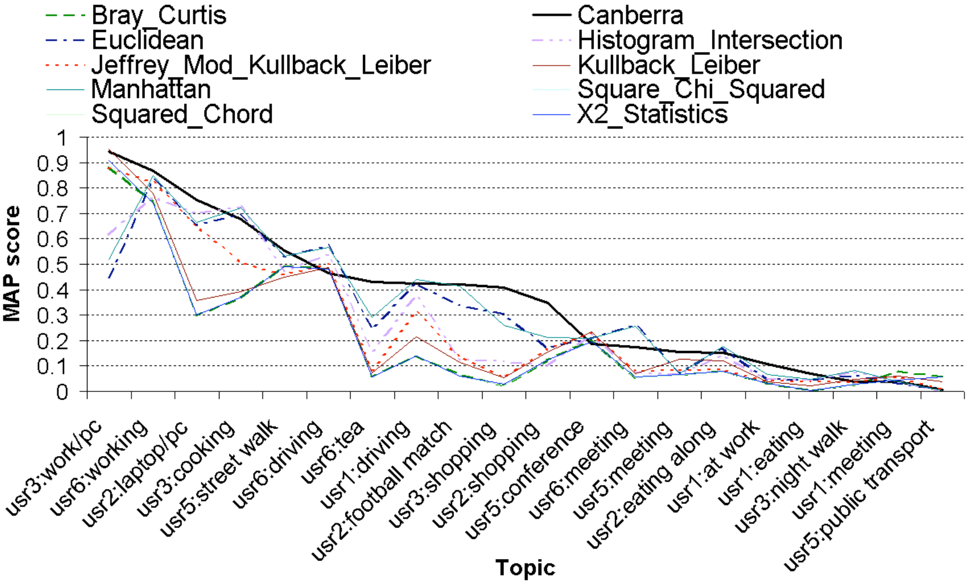


Figure 4.10: An examination on the retrieval performance affects on the ”general queries” test set by changing the vector distance method. Results are sorted by the Canberra method.

should be discarded (0.3702 MAP vs. 0.3695). There was quite a variance in the overall MAP score with changes in the vector distance metric used to compare MPEG-7 features (i.e. ranging from 0.3606 MAP for *Canberra* to 0.2123 for *Squared Chord* distance). As can be viewed in Figure 4.10, Figure 4.11, and Table 4.9 the *Canberra* distance almost always performs better than the other distance metrics. An explanation for why approaches such as *Canberra*, *Manhattan*, and *Euclidean* work well is probably linked to the fact that the scalable colour bin values can be negative as well as positive (The other 3 MPEG7 descriptors always produce positive bin values). The 3 mentioned vector distance metrics consider absolute differences and have positive denominators too, so therefore the effect of large negative values does not affect the final result for each comparison, which will always be positive. The other vector distance metrics can produce both positive and negative results and absolute magnitude values aren’t considered. This may be an explanation for the affected performance.

Approach	MAP
Canberra	0.3607
Manhattan	0.3209
Euclidean	0.3054
Histogram Intersection	0.2705
Jeffrey Mod Kullback Leiber	0.2613
Kullback Leiber	0.2391
Bray Curtis	0.2138
Square Chi Squared	0.2130
X^2 Statistics	0.2130
Squared Chord	0.2123
Pearson	0.1107

Table 4.9: This table lists the overall retrieval performance of each vector distance approach on the “general queries” test set.

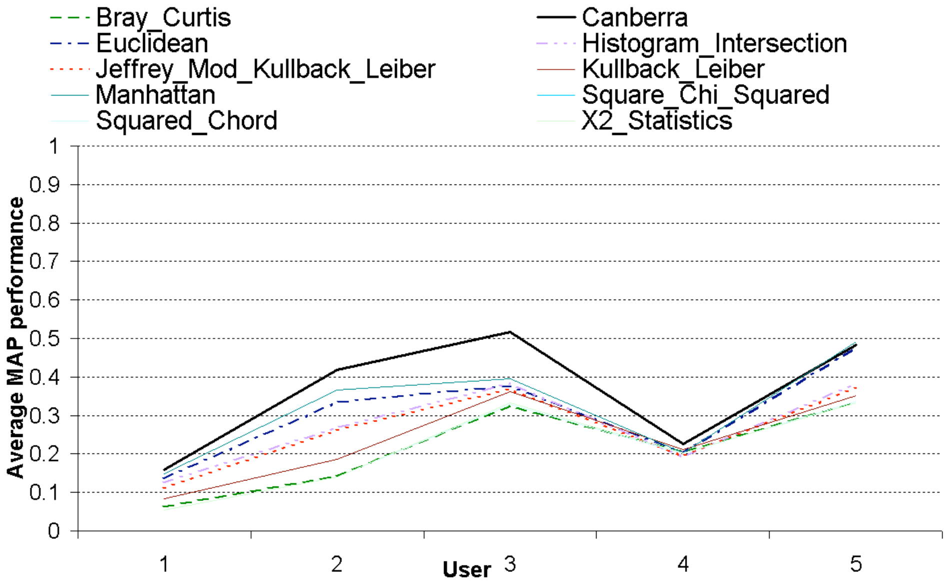


Figure 4.11: An examination on the retrieval performance affects on each user’s data in the ”general queries” test set by changing the vector distance method.

4.3.4 Optimal Normalisation Technique

Each and every data source will produce a score for how similar every event is to a given event. It is useful to combine these scores together to give an indication of how similar each event is to a given event, by considering all sources of information. However, before these sources of information can be combined together, it is firstly necessary to normalise these sets of scores to a common range. We now discuss the optimal normalisation technique that we have found.

The best trained normalisation technique was the *Min-Max* approach. This technique also worked best on the test set too as evidenced in Figure 4.12. In 17 out of 20 topics this technique was the best performing normalisation technique, especially when compared to the *Mean Shift* and *Sum* normalisation techniques. It is noted that the *Standard* approach (i.e. divide all scores by maximum score) provides the exact same scores as the *Min-Max* technique, therefore indicating many scores of zero were encountered in normalising the data sources before fusion[3].

4.3.5 Optimal Fusion Technique

Given a normalised score for each and every data source, it is then possible to combine these scores together. The advantage in doing this is that while the temperature source of information tells us that 2 sources of information are similar, the light and MPEG-7 sources may say otherwise. Only considering candidates where all sources of information indicate similarity will lead to improved retrieval performance, and leave an end user more satisfied with search results. Therefore we investigate: 1) Which sources of information should be considered, 2) How much confidence should be attached to each source, and 3) Which fusion technique to use to combine the sources together.

The best trained fusion technique was that of *CombMAX* (selecting the maximum normalised value from all the 5 sources). On the test set this was also the case with the

[3] The reason for this is the query event was compared to itself, thus always giving a similarity score of zero (the optimal score) in the set of figures. This would be dealt with in a post-processing stage of the application which removes this query event from the results list. This last pass occurs after the normalisation step however.

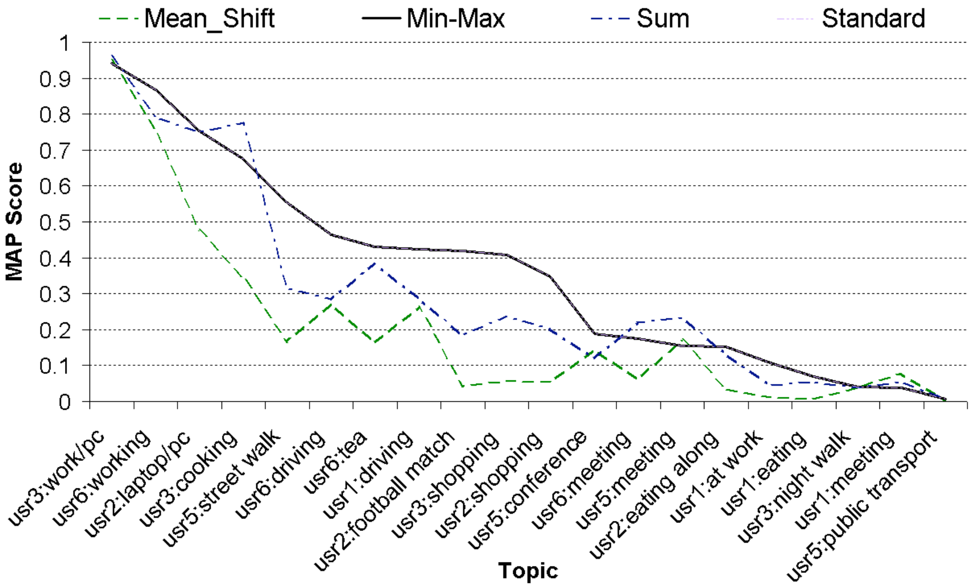


Figure 4.12: An examination on the retrieval performance affects on the ”general queries” test set by changing the normalisation technique.

CombMAX performing best, with strong performances also produced by the *CombANZ*, *CombSUM*, and *CombMNZ* approaches (Figure 4.13). The *CombMED* and *CombMIN* approaches perform quite poorly. These methods select the source with the smallest (or one of the median) scores. Its interesting to note that fused scores nearer zero indicate those events that are more likely to be similar to the reference event. The performances of *CombMIN* as opposed to that of *CombMAX* indicate that the optimal source to select is the *least* certain source. The more conservative *CombMAX* method performs best in this case, as it selects the least certain source. This mirrors the event segmentation experiments where the most conservative fusion technique (*CombMIN* in that case, as scores towards 1 were more desirable) proved most successful.

In terms of how much confidence to attach to each source of information, the best trained weights for the different sources on the training set are outlined in Table 4.10. As can be seen from Figure 4.14 the MPEG 7 source of information performs very strongly as more emphasis is put on it, whereas the overall performance suffers if too much emphasis is put on any one of the other sources (see Figure 4.15, Figure 4.16, Figure 4.17, and

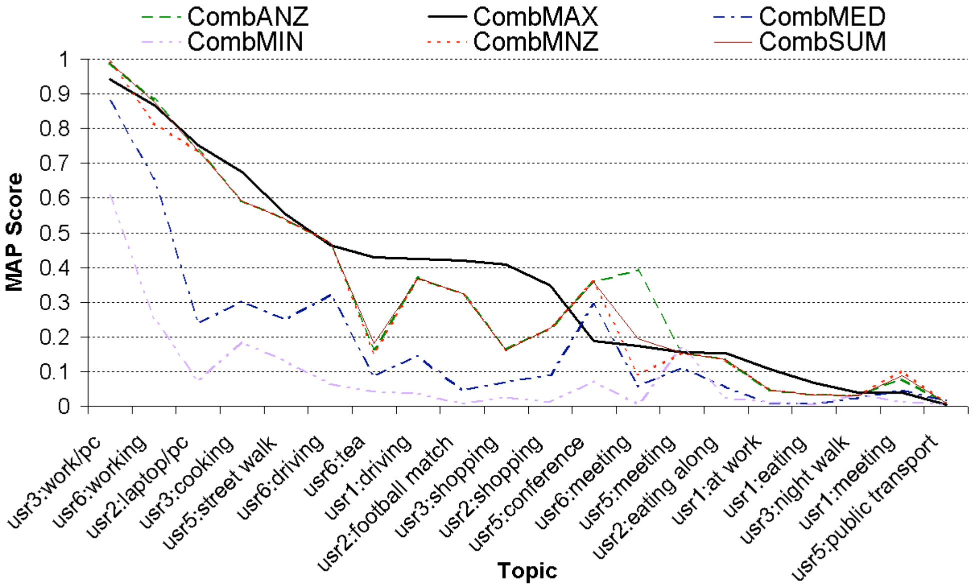


Figure 4.13: An examination on the retrieval performance affects on the ”general queries” test set by changing the data fusion method.

MPEG 7	0.293
Accelerometer	0.055
Light	0.203
PIR	0.197
Temperature	0.250

Table 4.10: A summary of the optimal fusion weights as determined in the “general queries” training dataset.

Figure 4.18).

Figure 4.19 displays the performance of the sources of information on their own. Again it can be seen that the MPEG7 data source is the most valuable single source in determining similar events.

4.3.6 MPEG-7 vs. Sensor Sources

Given that it can take up to 30 minutes to extract the four MPEG-7 descriptors for a day’s worth of images in the “general” dataset, it was decided to investigate the retrieval performance using sensor sources only. 84 different combinations of approaches were investigated in training over a time period of 74 minutes, and the best trained parameters using

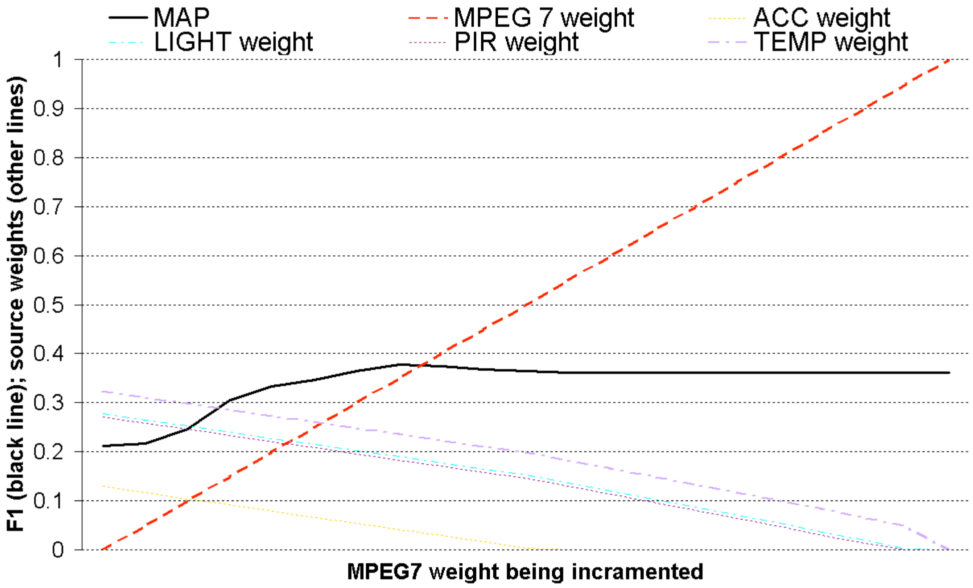


Figure 4.14: An examination on the retrieval performance affects on the ”general queries” test set by increasing the emphasis on the MPEG-7 Source.

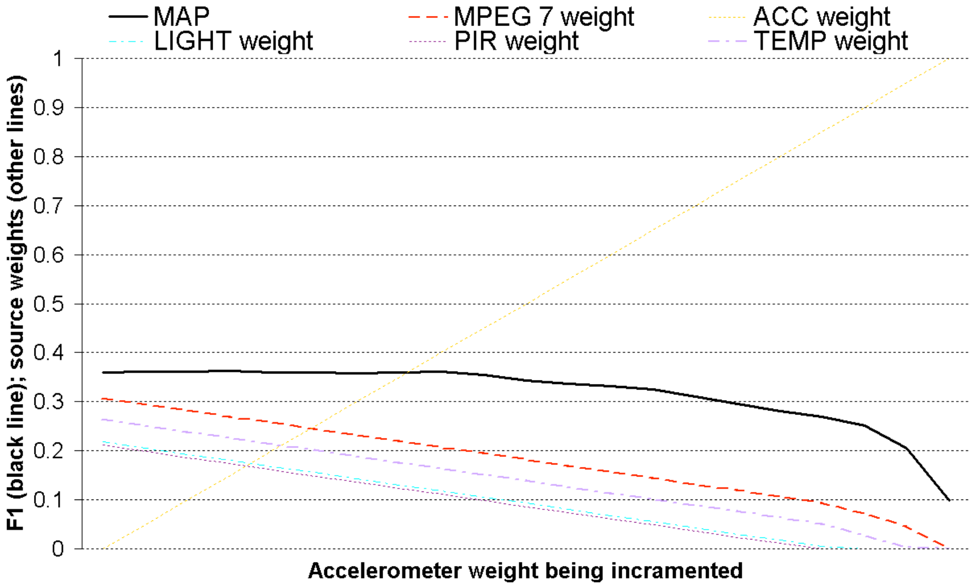


Figure 4.15: An examination on the retrieval performance affects on the ”general queries” test set by increasing the emphasis on the accelerometer source.

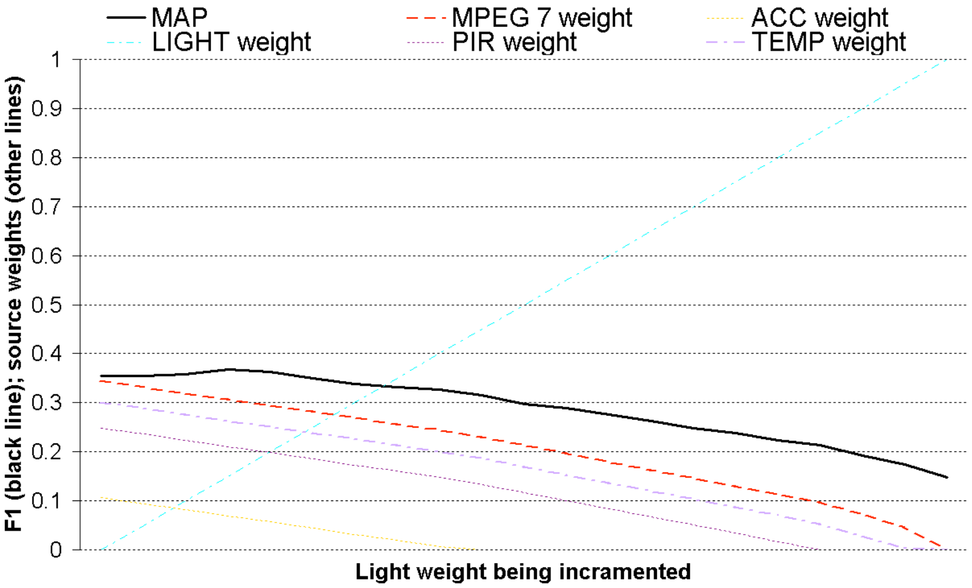


Figure 4.16: An examination on the retrieval performance affects on the ”general queries” test set by increasing the emphasis on the light source.

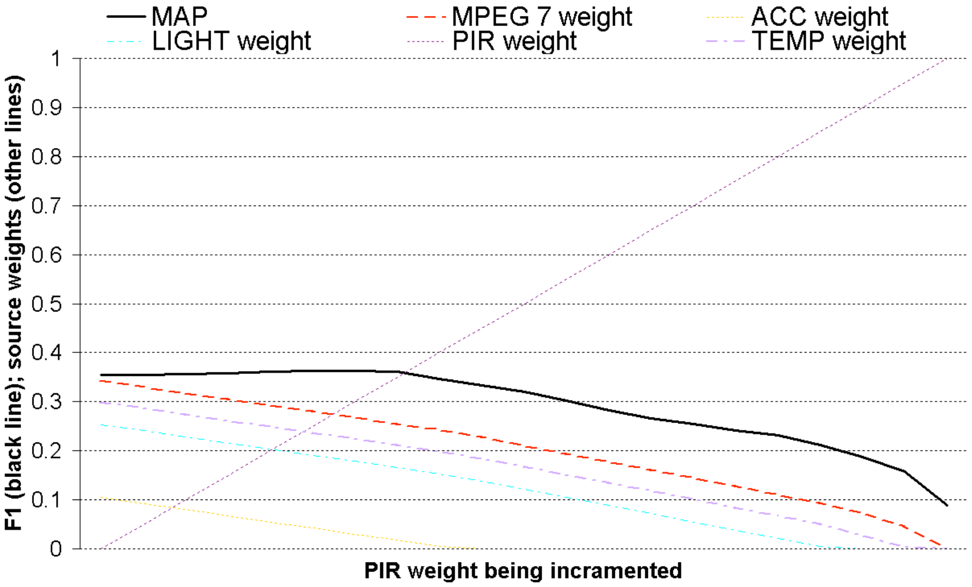


Figure 4.17: An examination on the retrieval performance affects on the ”general queries” test set by increasing the emphasis on the PIR source.

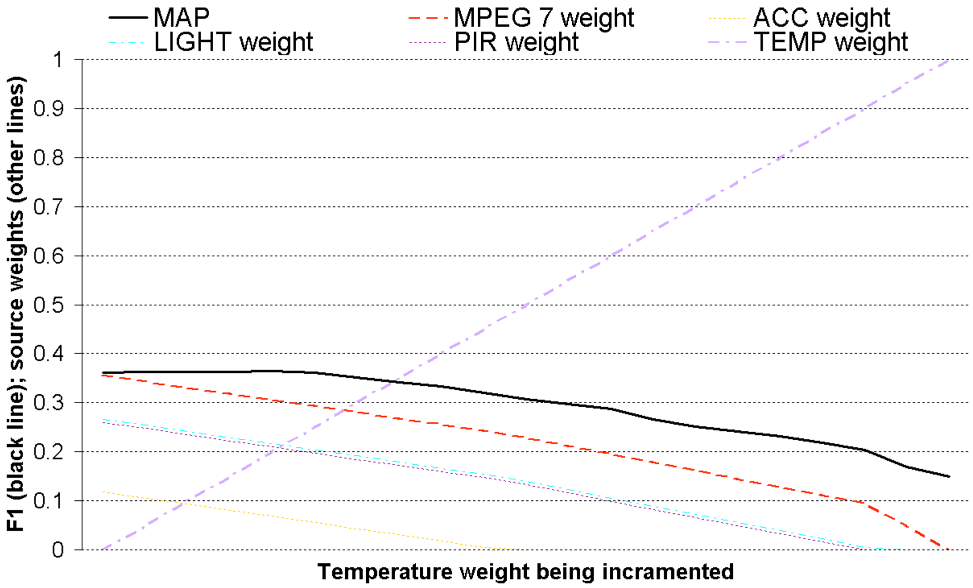


Figure 4.18: An examination on the retrieval performance affects on the ”general queries” test set by increasing the emphasis on the temperature source.

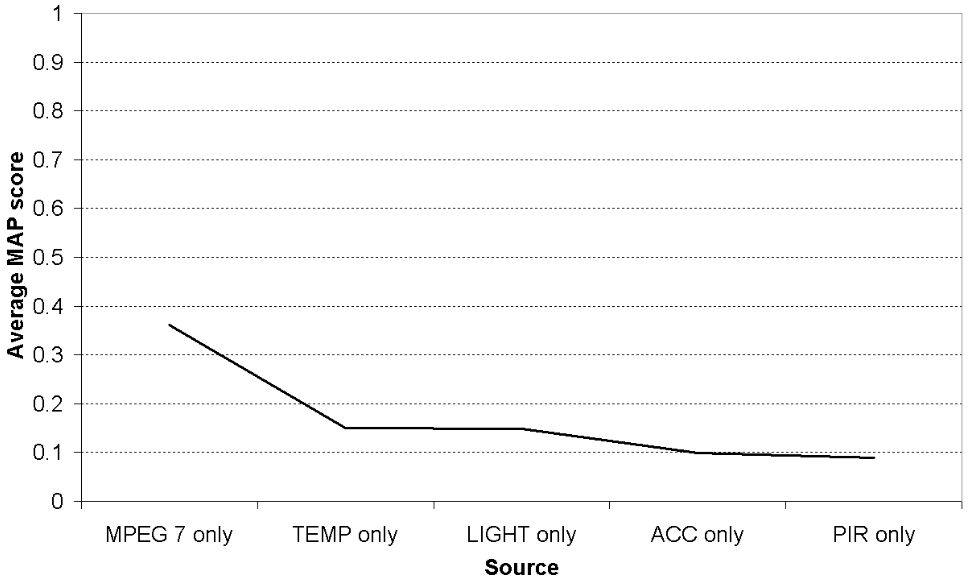


Figure 4.19: An examination on the retrieval performance affects on the ”general queries” test set when each source is applied individually and not fused with the other sources.

Research Variable	**Parameter Value**
Keyframe selection	event average
Vector distance	Canberra
Weight middle images	No
Normalisation	Min Max
Fusion	**CombANZ**
Within/cross event weight	N/A
MPEG7 features	colour structure, scalable colour, edge histogram
MPEG7 weight	**0.000**
Accelerometer weight	**0.220**
Light weight	**0.287**
PIR weight	**0.243**
Temperature weight	**0.250**

Table 4.11: A summary of the best trained parameters by using only sensor sources of information.

sensor sources only is outlined in Table 4.11. Bolded items indicate changes from the best trained system which includes MPEG-7 as well as sensor sources.

However on inspection of Figure 4.20 it is apparent that on only 1 of the 20 testing topics does the use of only sensor sources outperform the use of *all* sources of information. It is illustrated that the incorporation of the MPEG 7 source of information boosts retrieval performance across all topics. The inclusion of the MPEG-7 source of information is critical in terms of the precision @5 and precision @10 results. These metrics are very important in terms of returning relevant results early to the user, and thus improving the user experience of any system. For every given metric the inclusion of the MPEG-7 source of information improves performance.

It is interesting to note that taking only a small selection of the images from around the middle of the event can perform *almost* as good as considering all the images from the given event. Figure 4.21 illustrates how considering more images around the middle of an event (as we move right on the x-axis) leads to a performance improvement. However, in the training set if we just extract MPEG-7 features from the middle 35 images in an event, we will perform 89% as well as when all the images in the event are processed. However, this is very acceptable if we consider that instead of processing each and every image in the SenseCam collection, we can process just 34% of that collection, and still achieve 89% of

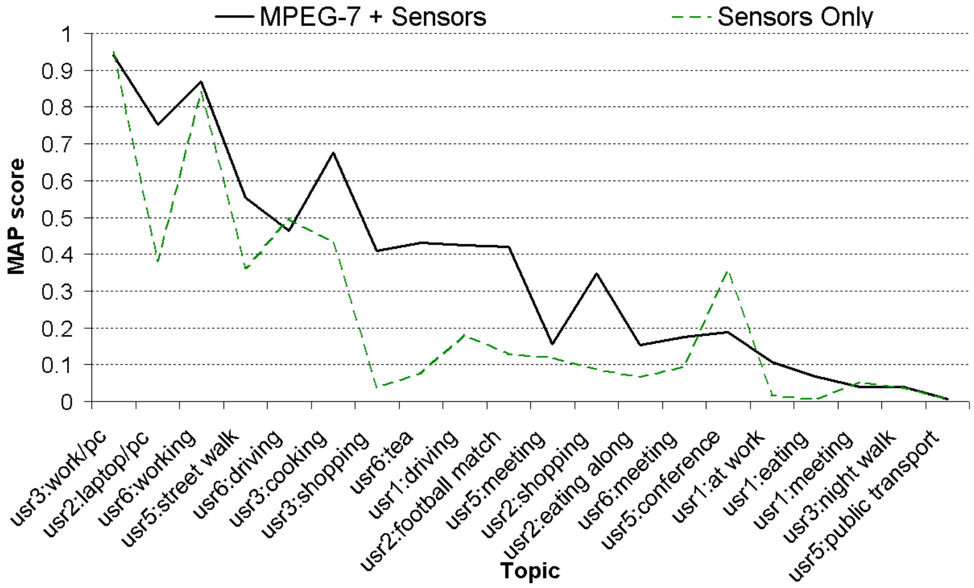


Figure 4.20: An examination on the retrieval performance affects on the ”general queries” test set by changing whether the MPEG-7 source of information should be included or not in the fusion process.

the retrieval performance.

This approach of selecting the MPEG-7 features of the middle 35 images in an event, in combination with using sensor sources, performs much better than considering only the sensor sources, as illustrated in Figure 4.22. While we need visual features for acceptable event retrieval performance, it is not necessary to extract features from each and every image, as after event segmentation (performed on only sensor sources of information), we can get good retrieval performance by just extracting the middle 35 images from each SenseCam event. This takes only 34% of the processing time of what it would take to extract features from *all* the images, or to express in another way, in our “specific queries” dataset of 1.86 million images, 1.23 million will *not* have to be processed.

4.3.7 Is Intelligent Segmentation Helping Retrieval Performance?

The previous chapter detailed intelligent multimodal techniques to segment sequences of SenseCam images into distinct events or activities. The aim was to detect semantic events

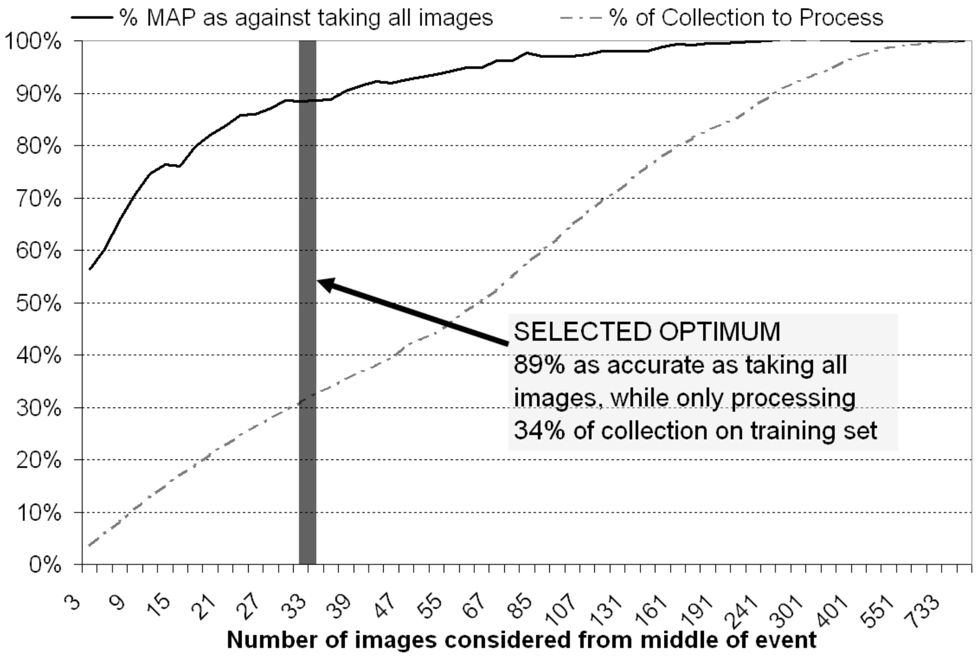


Figure 4.21: An examination on the "general queries" test set retrieval performance in processing only a selection of the images to represent an event.

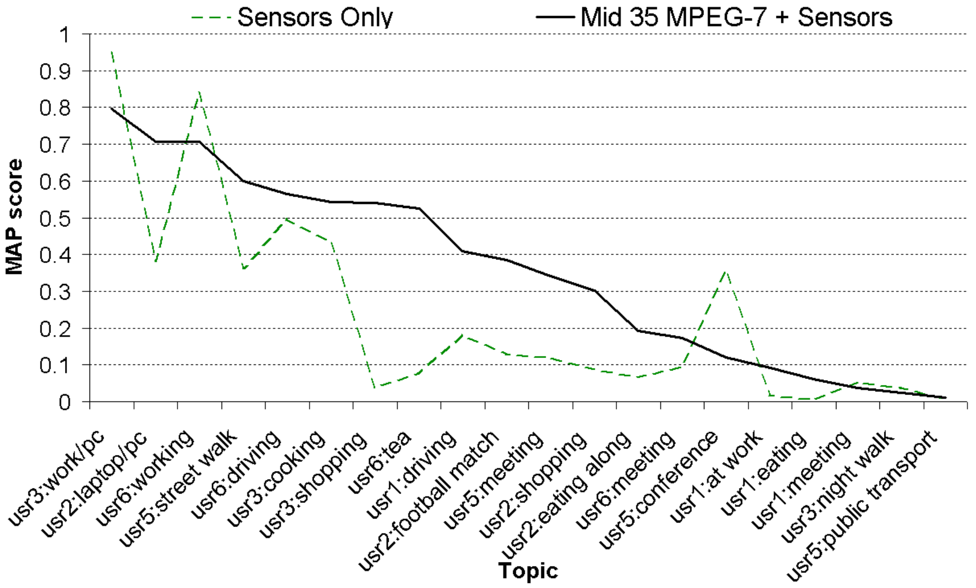


Figure 4.22: Processing middle 35 images vs. sensor sources only.

as close as possible to a user defined groundtruth. While this may be somewhat useful to the user while browsing through their images, as events are mapped as close as possible to the semantic user groundtruth, we also believe that another reason for intelligent multimodal segmentation may be that it aids retrieval, as similar images are grouped together which makes representing an event (as a unit for retrieval) easier.

To investigate the effects of good event segmentation on the retrieval performance two different systems were compared to each other:

1. The optimal multimodal segmentation system on which all results in this chapter thus far have been based on

2. A simple temporal segmentation system where no more than 95 images can exist in an event

All the data is available to investigate the retrieval performance based on the optimal segmentation system, in terms of images being segmented into events, in terms of the groundtruth being gathered, etc. Given that in the groundtruth of the segmentation experiments chapter, each event, on average, consists of 95 images; it was decided to divide the images of each user into events, each of 95 images in length, within available chunks[4].

Given that the events from both approaches will differ in terms of length (i.e. start times may be slightly different, and event durations can be quite different too), a decision was required on whether to construct a new groundtruth or to make an approximation of the existing groundtruth. Given the intensive annotation burden the former step would require, and also given that it is not the core focus of this chapter, it was decided to opt for the latter approach. To achieve this, the newly created events from the simple temporal segmentation application were aligned with the events from the original intelligent multimodal segmentation application, based on how much time overlap exists between events.

391 different approaches were investigated in training, and the tuned parameters using the simple temporal based segmentation are outlined in Table 4.12. Bolded items indicate

[4] A chunk is when all images are continuously captured, until a 2 hour break exists between adjacent images e.g. indicating when the SenseCam has been turned off for the user to go to sleep, etc.

Research Variable	Parameter Value
Keyframe selection	event average
Vector distance	Canberra
Weight middle images	No
Normalisation	Min Max
Fusion	CombMAX
Within/cross event weight	N/A
MPEG7 features	colour structure, scalable colour, edge histogram
MPEG7 weight	**0.256**
Accelerometer weight	**0.114**
Light weight	**0.213**
PIR weight	**0.150**
Temperature weight	**0.267**

Table 4.12: A summary of the best trained parameters using simple temporal segmentation.

changes from the best trained system using the intelligent multimodal segmentation approach (Table 4.7). As can be observed, the parameters of this best trained system on the temporally segmented events is very close to the best trained system on the multimodal intelligently segmented events, with only a slight change in weights the various sources. On the training set this system had a MAP score of 0.2533.

Of course it is of more importance on how well the temporal segmented events retrieval performance compares to the multimodal segmented events on the test set. In terms of the overall MAP score, the temporal segmentation system has a retrieval performance of 0.2700 in comparison to a MAP score of 0.3608 for the multimodal segmentation system. Figure 4.23 illustrates that for almost all topics the multimodal approach performs better than the simple segmentation approach, while Figure 4.24 illustrates that the average performance for every user is much better with the multimodal segmentation approach as opposed to the simple segmentation approach.

On close inspection of Figure 4.23 it is noticeable that the topics where there is the greatest disparity in the performance of both approaches are those topics such as “working on pc”, “at work”, and “driving”. By nature these events can be quite significant in terms of time duration, and with events being segmented every 95 images (approximately every

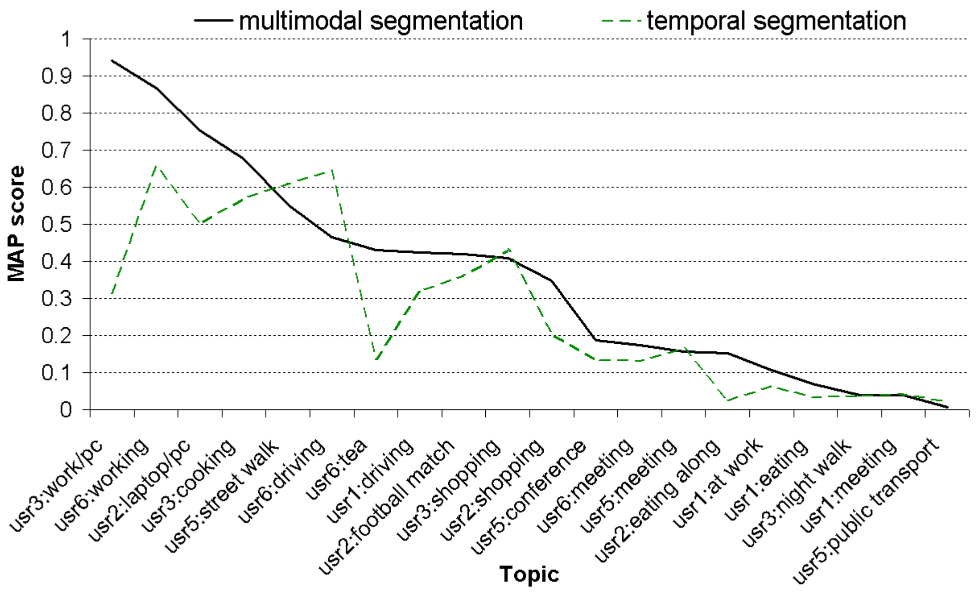


Figure 4.23: An examination on the retrieval performance affects on the ”general queries” test set by changing whether multimodal or temporal segmentation should be used.

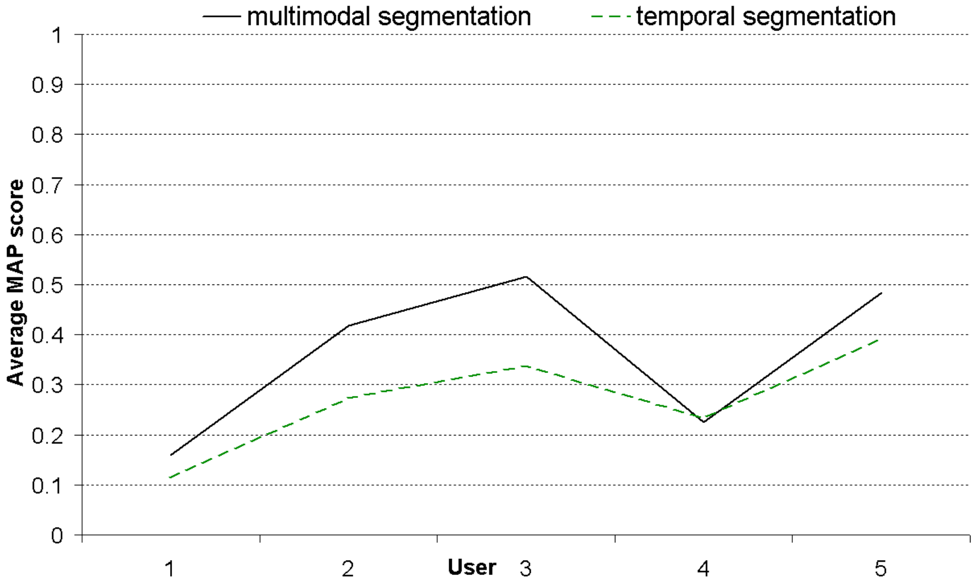


Figure 4.24: An examination on the retrieval performance affects on each user’s data in the ”general queries” test set by changing whether multimodal or temporal segmentation should be used.

Segmentation Technique	Multimodal	Temporal
Mean Average Precision	0.3608	0.2700
precision	0.1748	0.1591
recall	0.7021	0.6839
F-Measure	0.2798	0.2581
p @ 5	**0.6000**	**0.4100**
p @ 10	**0.5000**	**0.4150**
p @ 20	0.4375	0.3500
p @ 30	0.3983	0.3367
p @ 40	0.3725	0.3075
p @ 50	0.3480	0.2850

Table 4.13: A summary of multimodal vs. temporal segmentation affects on retrieval performance.

30 minutes), it is indeed possible that those events are under-segmented[5]. The multimodal segmentation approach also leads to much better retrieval performance in terms of P@5 and P@10 in particular (Table 4.13).

4.3.8 Retrieval Performance on User-Generated Specific Queries

Earlier in the chapter we introduced the datasets used in these experiments. To somehow capture the broad variety of information needs that users will have, we constructed two datasets. One with 16 quite "general" queries e.g. driving, eating, working, etc. The other dataset contained 23 user generated queries containing a very "specific" information need e.g. talking to Lynda, special dinners, on the aeroplane, etc.

While retrieval performance is quite good with respect to the "general" queries (e.g. P@5 for *event average* approach = 0.69 overall), it is challenging to retrieve relevant results for queries seeking a very "specific" information need (where P@5 for *event average* approach = 0.30 overall). Firstly this is most likely due to the fact that there are less relevant potential results in the users' lifelog for the "specific" queries, as the performance of the "general" queries are 158% better for P@5 across all the approaches, but then become 207% better for P@10, and then 235% better for P@20. This may indicate that there may

[5] We must stress that the temporal segmentation results were taken against a groundtruth inferred from the multimodal segmented groundtruth. Therefore these results can not be claimed to be conclusive.

be less than 10 potentially relevant items in the lifelog, if even that, for many of the users' "specific" queries. Another reason for the relatively poor performance of the "specific" queries is that many of them may require an associated semantic meaning.

4.3.9 Comparison to Other Techniques

In this chapter we have investigated techniques to retrieve similar lifelog events to any given event. Wang and his colleagues in Princeton perform event retrieval on a video lifelog but their approach is quite computationally expensive, in that when they compare two events, all images within the events are compared against each other [155]. Also they use audio features, and in prior work we have found that people are concerned about the recording of audio [50]. Therefore we compare one of our proposed techniques to two other retrieval techniques based on extracting interest point features from images e.g. identifying the corners of a computer monitor, so that other events can be retrieved even if the wearer was sitting in a different position than normal. Here we investigate the performance of retrieving events in a lifelog, of our "global" MPEG7-based image descriptor to the "local" SIFT and SURF image descriptors, which we will now describe below.

4.3.9.1 Approaches being evaluated

We will now describe the approaches being compared, a "global" MPEG-7 based approach that has been evaluated in earlier subsections, SIFT and SURF "local" image descriptors, and finally a fusion of our approach with a SURF approach to evaluate how complimentary both are.

- **MPEG7Sense:** This refers to the optimal system discussed in the previous subsections. The only difference is that events are represented by their *middle image* instead of the average feature vector of all the images in the event, which means that in terms of "general" queries there is a decrease of 0.364 vs. 0.188 in terms of MAP score. Given the computational complexity of SIFT and SURF, it was decided to only process the middle image of all the events. In order to allow a fair comparison of the

Research Variable	**Parameter Value**
Keyframe selection	middle image
Vector distance	Canberra
Normalisation	Min Max
Fusion	Comb MAX
MPEG7 features	scalable colour, edge histogram
MPEG7 weight	0.293
Accelerometer weight	0.057
Light weight	0.203
PIR weight	0.197
Temperature weight	0.25

Table 4.14: An overview of the best trained *"MPEG7Sense"* parameters that will be used in the "specific queries" dataset.

approaches, we thus make use of the *middle image* keyframe approach as opposed to the *event average* approach. Table 4.14 outlines this best trained system after over one thousand parameter variations were evaluated. This approach will be referred to as "*MPEG7Sense*" for the remainder of this section.

- **SIFT:** This is a method for extracting interest point features from images [92]. It detects interest point locations and also extracts features from around the points that can be used to perform reliable matching between different views of an object or scene. The SIFT features are invariant to image orientation, image scale, and provide robust matching across a substantial range of affine distortions, changes in 3D viewpoint, addition of noise, and changes in illumination. In addition to these properties, they are distinctive, easy to extract, allow for correct object identification with low probability of mismatch and are easy to match against a large database of local features. They are also robust to occlusion; as few as three SIFT features from an object are enough to compute its location and pose. In addition to object recognition, the SIFT features can be used for matching, which is useful for tracking and 3D scene reconstruction.

  For event/keyframe matching we use the approach described by Blighe *et. al.* [19] whereby SIFT features are first extracted from a set of reference keyframes and stored

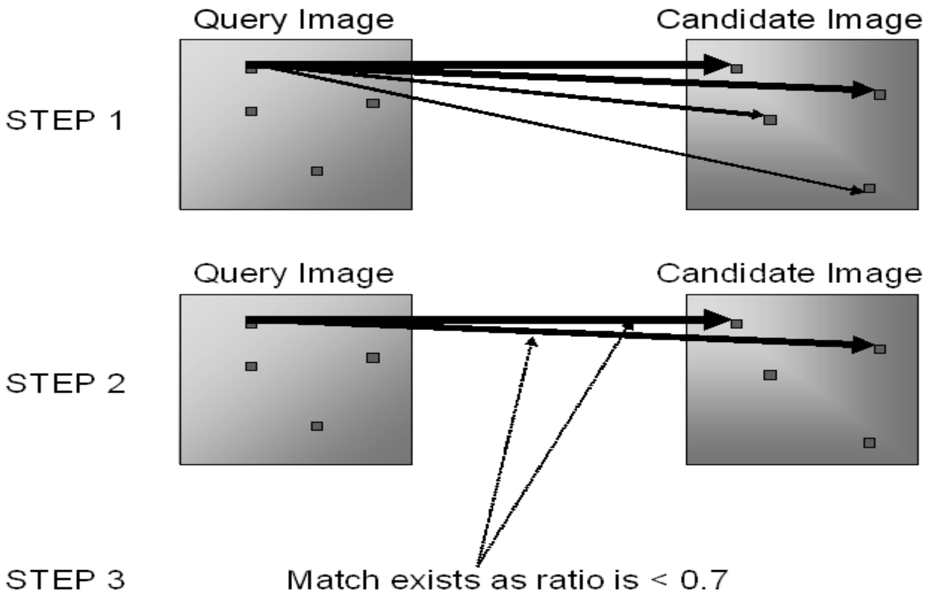


Figure 4.25: An explanation of the SIFT distance ratio test.

in a database. A new keyframe is matched by comparing each feature from the new keyframe to this previous database and finding candidate matching features based on Euclidean distance of their feature vectors. In order to match features between keyframes, the distance ratio test was used [19, 18]. To examine whether a point from the 1^{st} keyframe has a match in the 2^{nd}, it's two most similar descriptors in the 2^{nd} keyframe are found. If the ratio of the nearest distance to the second nearest distance is less than 0.7, a match is declared, as illustrated in Figure 4.25. The number of matches between a keyframe and all other keyframes in the event are summed, and then the average number of matches is calculated. Events are ranked based on the average number of matches calculated, with the most relevant items containing the highest number of matches.

- **SURF:** Introduced by Bay *et. al.*, Speeded Up Robust Features (SURF) are inspired by the SIFT feature approach, but speed up the extraction and description of interest points by exploiting integral images, achieving state-of-the-art performance in feature matching [14] .

While the extraction of SURF features is faster than SIFT, it is still computationally expensive to perform exhaustive matching between a query image and an image collection. Targeting fast image retrieval, Nistér and Stewénius describe an approach to image matching that scales to very large datasets, up to a million images [111]. Following their proposed scheme, we created a hierarchical visual word vocabulary using seven million SURF descriptors extracted from a collection of web images. These descriptors were clustered hierarchically using K-means, to generate a vocabulary tree with 4096 leaf nodes (visual words). An image database was structured by including an inverted file for each visual word, allowing efficient retrieval. Database images were compared to the query image using the Manhattan distance measure between their normalised histograms of visual words. This measure performed best in the original work [111]. Unlike the SIFT approach just described, which exhaustively compares every image to the query, a vocabulary tree query can be run in a few seconds for a database of thousands of images.

To further improve performance, the top 20 results are reranked by counting bi-directional matches between the query and each database image. In Lowe's work with SIFT features [92], a match between interest points was determined by using the distance ratio test. Since this measure is asymmetrical, we also compute the matches in the reverse direction (from the target to the query) and we count the matches that occur in both directions (bi-directional matches) as illustrated in Figure 4.26. Such matches were found to be very stable and strong indicators of a good match. The value of α was optimised using a training set and found that a value of $\alpha = 0.7$ worked best. This is to be expected, as SenseCam images are generally of a lower visual quality than the images used by Lowe.

- **Fused Approaches** As the global colour and edge features of MPEG-7 are complementary to the local SURF features, fusing the results of both approaches has the potential to improve the overall event matching performance. We describe here the two main approaches to fusion we investigated.

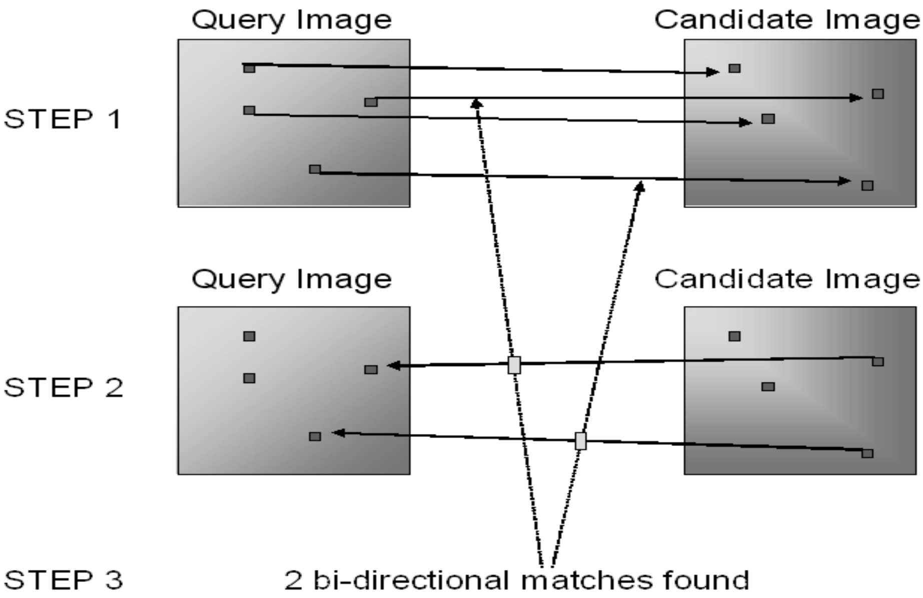


Figure 4.26: An explanation of the SURF bi-directional matches approach.

1. **Baseline Fusion** The bi-directional SURF matches that are used for SURF reranking are strong indicators of matching confidence. As a baseline fusion scheme, we take the *most confident* SURF results and then append the rank-list provided by the *MPEG7Sense* approach. More formally, we choose the top events ranked by the SURF scheme that have at least T bi-directional matches with the query image, then we insert the *MPEG7Sense* results that have not already occurred in our fused rank-list. We trained the parameter T on a training set of 24 queries and found that $T = 2$ optimised the MAP score.

2. **Score Fusion** In this approach we rank results in the following order for a given query based on: 1) events proposed by both *MPEG7Sense* and SURF, 2) events proposed by SURF, and 3) events proposed by *MPEG7Sense*. Parts 2 & 3 have already been ranked, so they are straightforward to order. However, for part 1 the scores from both sources must be normalised and fused. After training we selected the *Min-Max* normalisation technique and the *CombSUM* fusion method. A weighting/confidence of 80% was placed on the SURF source, and 20% on the *MPEG7Sense* source.

4.3.9.2 Comments on Individual Performance of *MPEG7Sense*, SIFT, & SURF

In this chapter we have proposed a technique to retrieve relevant SenseCam events to a given query event. We are interested in comparing this technique to adaptations of techniques proposed by others in domains quite close to lifelogging. We compare against *SIFT* and also against *SURF* based retrieval algorithms to now evaluate the effectiveness of our *MPEG7Sense* technique.[6]

The *MPEG7Sense* approach worked slightly better overall on both the set of 16 general queries and on the set of 23 specific queries than either SIFT/SURF. On the set of general queries it was the best performing individual approach on 9 of the 16 queries (0.188 MAP vs. 0.177 for SURF vs. 0.133 for SIFT), and on the set of specific queries it was the best performing individual approach on 16 of the 23 queries (0.126 P@20 vs. 0.089 for SURF vs. 0.033 for SIFT).

Examining results on the set of "general" queries, all methods perform well on finding *driving events*, with a P@10 of 0.7 or greater. These events are common in the data, and the reliable detection of such events would allow these commonplace events to be removed from view, facilitating more efficient user browsing. SURF performs best overall with a MAP of 0.417 (SIFT: 0.4137, *MPEG7Sense*: 0.1121). For an *eating event* query, the retrieval was quite poor, with a maximum MAP score was 0.061 (*MPEG7Sense*). The query keyframe was an image of the wearer's living room (while the SenseCam wearer is eating), and many returned results are of the same location, though often not involving *eating*. The low performance can be attributed to the difficulty in determining the semantic meaning of a query. Reasonably good results are obtained on querying *shopping events*. While the relevant events are visually quite different, they are usually cluttered scenes with many colours and edges. The *MPEG7Sense* colour and edge descriptors generalise well, and so do the SURF visual words, probably because some visual words capture properties common to shopping events. MAP scores for this query are SURF: 0.2835, *MPEG7Sense*:

[6] It should be noted that our *MPEG7Sense* technique is not the optimal found previously in this chapter, but to allow a fair comparison events are represented by the middle image in an event, instead of the best technique which represents events via an average feature vector of all the values.

0.2059 and SIFT: 0.0335.

On the 23 "specific" topics it is interesting to consider Figure 4.27 which plots the P@20 score of the 3 approaches. Selecting 3 sample events where each approach works better than the other two, it is somewhat unsurprising that the *MPEG7Sense* approach works best on the "beaches" query, as returning back other events full of predominantly blue and yellow/orange are quite likely to be relevant, while there were not many features present in the given images explaining the relatively lower performance for SIFT and SURF. The SIFT approach works well on the "Lynda" query, where one user wished to compile all times he was talking to his colleague Lynda. These images were all taken in the same building where objects within the office would all have a similar look and feel. SIFT worked well in this case because the objects, and their spatial arrangement, within the images are visually similar to the query event. The SURF approach works well on the special dinners query. The SURF encoded feature points may detect items such as elegant glasses or salt cellars, and then by reranking the top results based on bi-directional matches, there may be a number of other dinner events excluded as they are not visually similar to the query event "special dinner".

4.3.9.3 Benefits of fusing *MPEG7Sense* & SURF

As we are evaluating the effectiveness of an adaptation of our approach (*MPEG7Sense*) against two other techniques (*SIFT, SURF*), we now investigate the benefits to retrieval of fusing these approaches. Given that the SIFT approach is computationally very expensive, we only evaluate fusions between the *MPEG7Sense* and *SURF* approaches. We now detail the benefits of fusing the results of both techniques together to provide the user with a greater number of relevant events in higher ranked positions.

Both the *baseline fusion* and the *score fusion* approaches offer improvements over individual runs in terms of MAP on the "general" set of queries. The best performing individual source on the set of general queries had a MAP score of 0.188 (*MPEG7Sense*). The *baseline fusion* approach had a MAP score of 0.203 and was at least as good as the best result

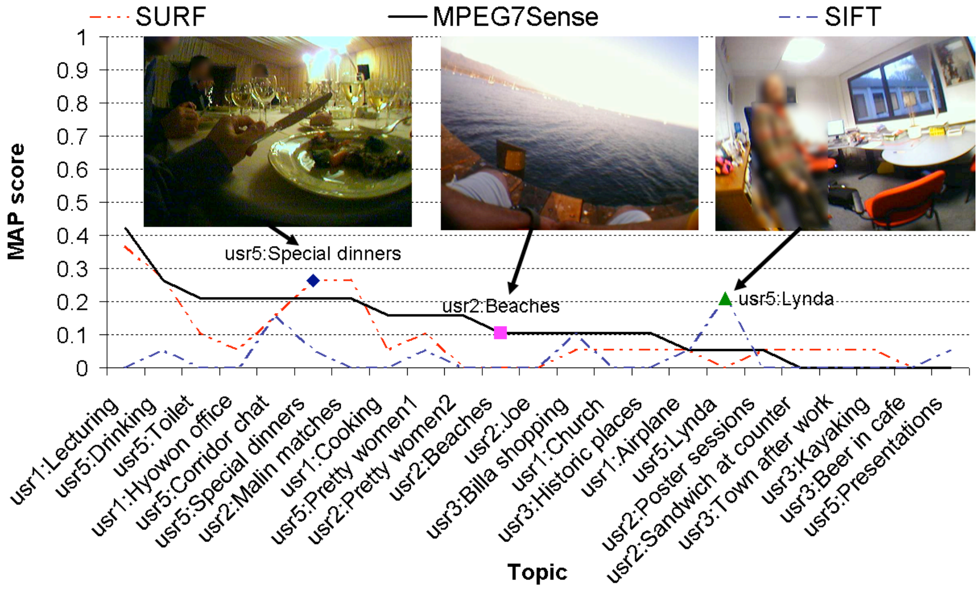


Figure 4.27: A summary of the P@20 performance of the *MPEG7Sense*, SIFT, and SURF approaches on the "specific queries" dataset.

offered by either the individual *MPEG7Sense* or SURF approaches on 7 out of 16 "general" topics. Meanwhile the *score fusion* approach had a MAP score of 0.201 and was also at least as good as the best results offered by either the individual *MPEG7Sense* or SURF approaches on 5 out of the 16 "general" topics.

The performance of both fusion techniques is perhaps even more effective when dealing with the difficult queries that have a very "specific" information need. The best performing individual source on the set of "specific" queries had a P@20 score of 0.126 (*MPEG7Sense*). The *baseline fusion* approach had a P@20 score of 0.130 and was at least as good as the best of the *MPEG7Sense* and SURF approaches on 16 of the 23 queries. Meanwhile the *score fusion* approach marginally outperforms the *baseline fusion* approach with a P@20 score of 0.135 and improves on the performance of any individual SURF or *MPEG7Sense* approach on 15 of the 23 "specific" queries.

It is interesting to note that the *score fusion* approach returns a greater number of relevant documents earlier than the *baseline fusion* approach on the "specific" queries as illustrated in Figure 4.28. This effect is also present on the set of "general" queries also. An

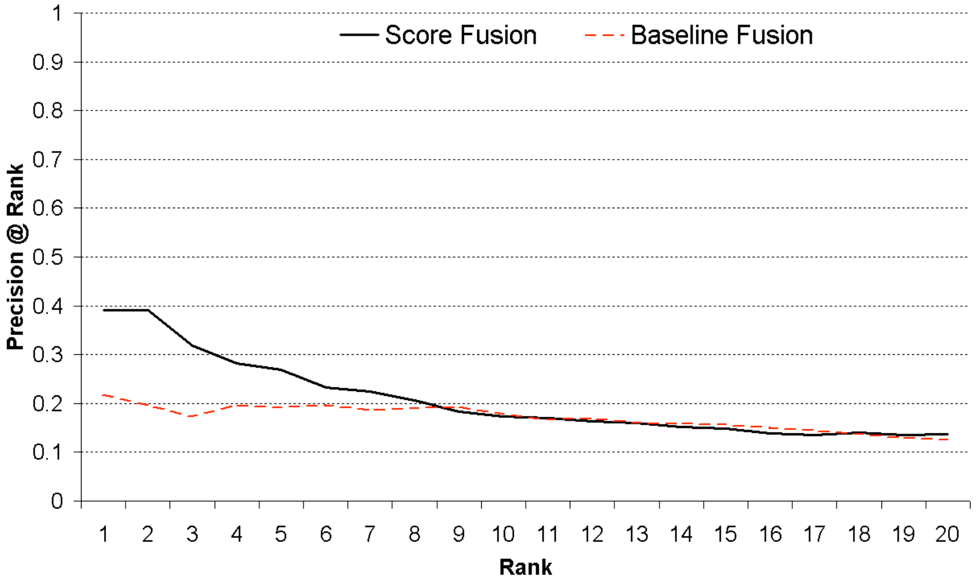


Figure 4.28: Comparing fusions at different "precision at" levels (x-axis) on the "specific queries" dataset.

explanation for the phenomenon is that the *score fusion* approach ranks candidates that are returned by both SURF and *MPEG7Sense*, and ranks those firstly. This may explain why for P@1, P@2 ... P@8 scores are all better for the *score fusion* approach than the *baseline fusion* approach. However, for candidates returned at ranks 10 and above the *baseline fusion* is the best performing approach. The likely explanation for this is that the *score fusion* strategy takes a more integrated approach to combine the retrieval scores, instead of the *baseline fusion* strategy of using SURF then switching to *MPEG7Sense* when the number of bi-directional matches drops below a threshold.

4.4 Summary

Considering that the human memory stores items associatively, in our hypothesis we have stated that we aim to *"... usefully provide the ability to link similar events together ... "*. The focus of this chapter was to identify the optimum technique to automatically compare lifelog events against each other. This is helpful for two reasons, to enable users to search for similar events to a given event (e.g. "show me other times where I was walking in the park"), and to help identify those events that frequently occur (e.g. a user may be in front of their PC constantly, which may indicate this is a routine/mundane event). This chapter is presented with significant research challenges owing firstly to the "semantic gap" [143], and secondly to the poor quality of SenseCam images which accentuates the challenge[7]. Lifelog events consist of many images, so the first challenge is in representing an event, and only thereafter can event representative values be compared against each other to indicate other events that may be similar to a given event. A number of techniques are investigated in this chapter: the optimal event representative selection technique; whether images towards the middle of the event should be weighted more strongly; the optimal vector distance technique to compare two images (or event representative vectors) against each other for the challenge of retrieval; and finally selection of the best normalisation and fusion techniques given the use of both content and contextual sources.

As described earlier we used two datasets, one of approximately 270,000 images with associated "general" queries (e.g. driving, at work, shopping, etc.), and another of approximately 1.8 million images with associated queries containing a very "specific" information need (e.g. special dinners, other times talking to Joe/Lynda, etc.). While performance is quite good with respect to the "general" queries (e.g. P@5 of *event average* approach = 0.69 overall), it is challenging to retrieve relevant results for queries seeking a very "specific" information need (where P@5 of *event average* approach = 0.30 overall). Firstly this is most likely due to the fact that there are less potentially relevant results in the users' lifelog for the "specific" queries, as the performance of the "general" queries are 158% better for P@5 across all the approaches, but then become 207% better for P@10, and then 235% better for P@20. This may indicate that there are less than 10 potentially relevant items in the lifelog, if even that, for many of the users' "specific" queries. Another reason for the poor performance of the "specific" queries is that many of them may require an associated semantic meaning, which is very difficult to extract given that by their nature these

[7] Given the prevalence of the SenseCam in the visual lifelogging domain, we have to accept that the images being used will be of a low resolution

queries had a very "specific" information need e.g. "show me events when I was talking to Lynda".

The best trained system achieved a MAP score of 0.3608 (with much deviation between topics as illustrated in Figure 4.7). The optimal way to represent an event of images is to construct an average value of all the MPEG-7 vector bins and contextual scalar sources of information; in fact the *Event Average* approach worked best in 15 out of the 20 test set topics.

It is very interesting to note that the approach towards modelling the event is influential in terms of retrieval performance. Applying standard techniques from the video retrieval domain, an obvious approach to represent an event is to select the *middle image*, however our *event average* approach didn't represent events by a keyframe image, instead representing events by taking the average feature vector value of all the images present in the event. This led to a 38% increase in retrieval performance on the set of "specific" queries, and an impressive increase of 106% on the set of "general" queries. However, as displayed in Figure 4.21, by only processing the middle 35 images of each event (just over 30% of the entire set of images), 90% of the retrieval performance of considering all the event's images is achieved. Indeed the approach of selecting the MPEG-7 features of the middle 35 images in an event, in combination with using sensor sources, performs much better than considering only the sensor sources, as illustrated in Figure 4.22. While processing on contextual sensor sources of information is very quick, we have illustrated the need to understand the image-based content information in terms of effective retrieval of events in a visual lifelog.

While it appeared sound to weight images towards the middle of an event more strongly, in practice this approach did not offer any performance gains. There was quite a variance in the overall MAP score with changes in the vector distance approach used to compare MPEG-7 features (ranging from 0.3606 for *Canberra* to 0.2133 for *Squared Chord* distance); the *Canberra* vector distance metric is recommended. The *Min-Max* normalisation approach works best while the *CombMAX* fusion approach is optimal. The MPEG-7 source of information is the most valuable data source in terms of retrieving similar lifelog events, in fact so much so that the introduction of this source in combination with the sensor sources results in a MAP score retrieval gain of 52.6%, and gains of 62.2% for the precision of the top 5 results, and 58.7% for the precision of the top 10 documents.

To illustrate the relationship between the current chapter and the previous chapter on event segmentation, we have indicated that intelligent multimodal event segmentation may lead to as much as a 33.6% boost in terms of retrieving similar events (over simple temporal segmentation as used by Wang *et. al.* [155]).

We have compared our approach against two distinct content based image retrieval methods: SIFT, & SURF. We have detailed that our approach works slightly better than the results from these two techniques. All three approaches are also highly complementary e.g. our approach works well on "beaches" type queries where dominant colours in the scene are important, SIFT adds many relevant results to queries of particular office settings, while SURF adds many relevant results on queries such as "special dinners". After training, we investigated two fusion techniques in this work, whereby we combined the results from our approach and the SURF approach. *Baseline fusion* approach performed slightly better overall, and we recommend use of this approach if the user wants a large number of relevant results. *Score fusion* approach performs optimally in returning relevant events in highly ranked positions (as illustrated in Figure 4.28), and as such this approach is recommended for users who do not want to sift through irrelevant results returned in highly ranked positions after querying.

4.5 Conclusions and Contributions

- While extracting image features is not necessary for event segmentation, it is vital towards acceptable event retrieval performance.

- How events are modelled has the most significant influence on event retrieval performance. The *Event Average* representation technique performs best. However we have discovered that by only considering up to the middle 35 images in an event, we can achieve 90% of the optimal retrieval accuracy while only having to process 30% of the content.

- While there is a variation in final system performance by changing the vector distance method, the simple *Manhattan* and *Euclidean* approaches perform almost as well as the optimal *Canberra* approach.
- The *Min-Max* normalisation technique is recommended while the *CombMAX* fusion technique is strongly recommended.
- Currently a query event is compared to *all* of the events in one's lifelog. A more scalable method to consider in future may be that of storing an inverted list of "visual words" as was implemented in the SURF lifelog retrieval approach in this chapter.
- We have published papers on event retrieval experiments in the 2008 MIR conference [47].

Chapter 5

DETERMINING MOST IMPORTANT EVENTS AND SELECTING KEYFRAME IMAGES

> *In order for visual lifelogging technologies to effectively provide autobiographical memory retrieval cues it is necessary to automatically:*
>
> ...
>
> - *accurately determine which of those events are most unique*
>
> ...

As the brain strongly encodes distinct memories, we wish to ease the burden on the users in finding such events from their lifelog through suggesting more interesting events to review. This chapter addresses the problem of trying to determine the most important events in an individual's day e.g. going to a rock concert is more significant than having your usual morning breakfast. Also in this chapter we address the challenge of selecting a given image from an event that is representative of that event, i.e. to provide the user with a summary image of an event (on average consisting 100 images), which one should be chosen? We will now introduce these challenges in further detail.

On average a user will have approximately 20 distinct events in a day (i.e. around 7,000 events per year depending on the user's lifestyle), but it is quite likely that many of these events may be routine or repeated events that are unimportant to the user. However, there will also be certain events that may be of great significance to the user, e.g. lunch with a loved one, an awards ceremony, attending a big football match, etc., because they are more unusual or novel. Thus it is desirable to automatically emphasise the events that may be more interesting to the user than the routine or mundane events.

Attempting to determine the "importance" or "significance" of an event is a very subjective exercise. Only the user can decide on the true importance of an event, e.g. *your* first meeting with a partner, the first time *your* child walked, when *you* received an award that was special to *you*, etc. The importance or significance of events can also change over time, e.g. meeting a stranger on the street (mildly important), may in fact be reflected on in the future as the first time when a couple met (very important). These very important/significant events occur quite sporadically, but in each and every day there are always some events that are at least more interesting to review than others. These events may be unique in that they are very different to everything else, or at least they are somewhat distinct or interesting in that users prefer to review these events. For simplicity of explanation throughout this thesis we refer to finding the "importance" of such events, whereby we mean that we are trying to highlight those events that the user is more likely to want to review (owing to the fact that such distinct activities may have been strongly encoded in his memory).

Figure 5.1 abstracts a user interface that utilises the notion of highlighting event importance. Such an interface will display various events from a given day, but those that appear larger in size are determined as being more important by an automated approach, whereas events with smaller sizes have been automatically determined to be mundane/routine events. This dissertation will now discuss various approaches taken to automatically determine the importance of each and every event in a user's lifelog. This is a considerable challenge but

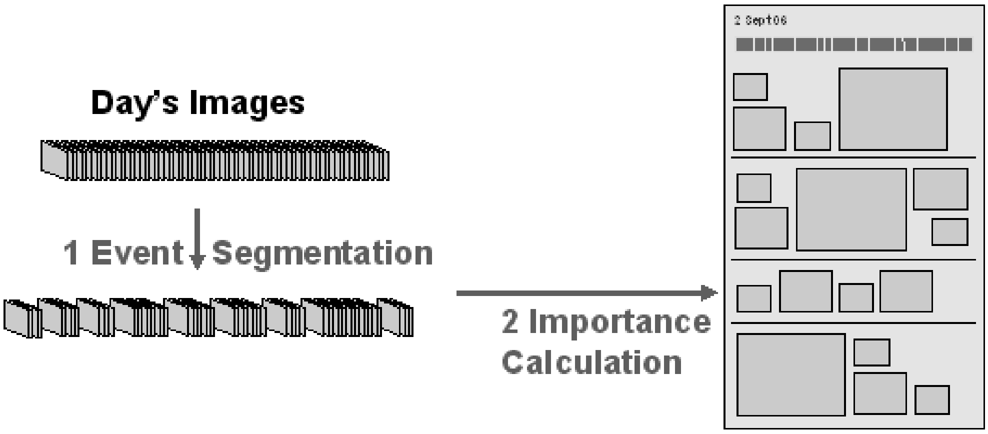


Figure 5.1: An overview of determining and then displaying lifelog events based on their importance.

potentially very rewarding in terms of assisting users to find those interesting/novel/important events that they are attempting to browse to. The various approaches taken to realise this goal will be explained in due course.

Given an interface like the one abstracted on the right hand side of Figure 5.1, it will be necessary to select a representative/keyframe image from each event, so as to help the user recognise what had happened in such an event. The use of keyframes is not unique to the domain of lifelogging as keyframes have been used in video retrieval as a means by which an at-a-glance summary can be offered to users. Keyframes are still images, which have been automatically extracted from video content. Digital video content is segmented into smaller units known as "shots", with a single keyframe used to represent each "shot" - this concept is somewhat similar to "events" within visual lifelogs. The frame(s) of video to be used as a keyframe is determined by attributes of the video content such as motion or the presence of faces. Cooper and Foote [41] note that *"keyframes must both accurately represent the underlying video clip and distinguishes that clip from the remainder of the collection"*. As such, an ideal keyframe summarises the major concepts contained within a media segment allowing a user to identify segments relevant to their information need. The selection of keyframes for visual lifelog content is however not without challenges unique to the domain. The keyframe selection methods, which have shown success and prevalence in the domain of video, may not translate directly to visual lifelogs, due to the lower rate of capture and the fisheye lens which makes SenseCam images distorted. In this chapter we will explore and evaluate various techniques that are traditional to the video domain, but also introduce and evaluate a new technique which shows much promise.

5.1 Approaches to Event Importance and Keyframe Selection

This section will now describe in more detail the approaches taken to realise the automated processing of these goals. We will firstly describe our approaches to determining the importance of each event, and thereafter we will discuss the various keyframe methods explored.

5.1.1 Event Importance Approaches

This subsection discusses a number of approaches to determine the importance of an event. There are two main categories, namely face detection and novelty. Prior work by Tancharoen & Aizawa introduced the concept of detecting faces to determine event importance, on the premise that they are a good indicator of face-to-face conversations, and hence social interaction which is quite important to many people [145]. We have introduced the concept of novelty detection, whereby using the visual and sensory properties of each and every event, we can calculate how unique/novel that event is. More unusual events may indicate that the user is doing something quite interesting. We will now discuss these approaches in detail.

5.1.1.1 Face Detection

Conversation detection is based on the presence of faces in SenseCam images, on the premise that when talking to someone we are facing them, thus this can be determined via the automatic detection of faces from SenseCam images. The current state of the art in this area of lifelogging is to detect the presence of faces to determine the importance of lifelog events [145]. We used the Intel OpenCV [23] face detection implementation. A number of face detection trained classifier models are provided with this library and other parameters can also be tuned such as the image scaling factor and the window size that a face should be present in. As SenseCam images are of low quality (generally 25kB each) and taken using a fisheye lens, we investigate which of these models is most suitable for lifelog images. The importance value for each event is determined by the number of images with faces in the given event (*relative to the total number of images in that actual event*). Higher scores indicate longer face-to-face conversations, and thus more importance should be attached to such events.

5.1.1.2 Novelty

Generally those events that humans cherish most are those that occur infrequently, e.g. going to a rock concert, talking to an old friend at the weekend, going to a restaurant for a meal in mid-week, etc. On the other hand routine or frequently occurring events do not give a user as much satisfaction while reviewing their day, e.g. being at work, driving into work, having breakfast, etc.

Based on this, within a day we determine those events that are more novel/unusual as opposed to those events that are routine/mundane. The approaches proposed in the lifelogging domain by this thesis is based on the concept of novelty, which has been used quite successfully in the text and video domains before [33, 58].

Finding the most novel event within a dataset is relatively straightforward, in that the event that is most different to all the other events in the set is selected. However, there are three main research questions: 1) What events to determine the novelty of? 2) What dataset to utilise in determining the novelty of those events? And 3) How to compare events against each other to determine a novelty score?

1. **What events to determine the novelty of?** Many multimedia and lifelogging systems digest large amounts of data into distinct days to then allow users select relevant days from a calendar [115, 81]. A day is a distinct and natural unit of information for a user to review. Therefore we will determine the novelty of each and every event in

a given day.

2. **What dataset to utilise in determining the novelty of those events?** To determine the novelty of an event within a day, what other events should it be compared against? Should just all the events of today be considered, or should events from previous days/months be considered too? If a large number of comparison events are chosen, the processing load may be quite high. However, if too few comparison events are chosen, the system may not be able to accurately model the user's typical behaviour.

3. **How to compare events against each other to determine the novelty score?** How do we compare two events against each other to somehow determine their similarity? This requires the comparison of pairs of events against each other, which was the focus of the previous chapter. To recap on the previous chapter's findings, this is realised through representing each event by the average value of all[1] the SenseCam images composing that event. These event representative features are then compared against each other using a combination of MPEG-7 visual features, accelerometer, light, passive infrared, and temperature sources.

Therefore given that it is aimed to determine the event that is most novel in a given day (1), and also given that events will be compared to each other using the optimal approach in the previous chapter (3), the primary research question is what dataset should be used to determine the novelty of the given events in a day (2)? Five separate novelty detection approaches which vary the "window" for novelty detection are proposed as follows:

Previous 7 Days To determine the importance of a given event in a day, we investigate how dissimilar each event is to all the other events of that day and of the previous 7 days. The premise of this approach is to take into account the range of activities occurring during the week, which should contain a number of routine/mundane events (break-

[1] It would also be fine to represent the event by the middle 35 images which performed very well in Chapter 4, but for this work we decided to go with the optimal technique in Chapter 4 of considering all images in the event.

fast, being at work, etc.) as well as identifying the outlier events such as a barbecue the wearer may have had on a Saturday afternoon.

Previous 3 Days This approach has half of the processing load of the *Previous 7 Days* approach and may highlight more strongly an important social event that occurred at home during the weekend, as it would be different to the previous 3 days of work. By considering the previous 7 days such an event may have not looked as important/unique as it would be quite similar to the events from the previous weekend.

Previous 14 Days Computationally this approach is more expensive, but a possible advantage will be to decrease the novelty of regular events after a brief break from an individual's routine. Consider an employee who works one week, then goes on holiday for a week, and then back to work the following Monday with a social gathering that evening. By considering the previous 14 days it can be seen that in the week prior to being on holiday, the employee spent a considerable amount of time in work, thus the social event of that Monday evening would then score as being more highly novel than the day at work (which may appear novel if only the previous 3 or 7 days are considered).

Events within ± 2 Hours from Previous 7 Days This is a more sophisticated approach that considers only events from the previous 7 days that are within ± 2 hours, on the 24 hour clock, of the event we are trying to compute the novelty of. The premise for this is to highlight, for instance, a family meal out in a nice restaurant at the evening time rather than the normal evening meal at home.

Previous 21 Days Constrained to Same Week Day In this approach only events occurring on the same day of the week over the past 3 weeks are considered e.g. A doctor may attend a meeting every Thursday, and when data from only the previous week is considered this may appear unique, however when data is taken from the previous 3 Thursdays, this may rightly appear as a routine event.

5.1.1.3 Combining Face Detection and Novelty

Considering the face/conversation detection approach, it is likely that there will still be many events which will have a face count score of zero yet which are novel! Therefore it is important to take these into consideration and rate the uniqueness of these events to determine which are least interesting. A result of this rationale is that another approach will be investigated whereby the importance of an event is ranked firstly by the concentration of images with faces present, and secondly ranked by the novelty score of that event, using the best performing novelty approach from the previous sub-section.

5.1.1.4 Post-Processing Step

After initial observations the relevant novelty approaches were returning too many events consisting of a large number of dark images as the most novel events of the particular days e.g. when the device is behind one's coat/jacket or in a very dark environment. The approaches appeared to detect the more frequently occurring events and mark them as mundane, however they selected too many toilet/under shirt/black events as being important. Technically these events are very dissimilar visually to all the other events which is made more acute given the strong weighting on the MPEG-7 source when comparing any two events against each other.

Therefore a post-processing step was introduced to remove the events that consist almost entirely of uniform, dark images. These can occur when the SenseCam is occluded behind the wearer's coat, is tucked under their shirt, or taken in a very dark environment where the images can not be recognised. The colour layout descriptor captures the spatial distribution of colour in an image. Therefore if the variance between the bins, taking into consideration all the images in a given event, of this descriptor is very low then it is most likely the event consists of many images of uniform colour (with the SenseCam this corresponds mainly to events of many very dark images). As a result of this an importance score of zero (the minimum score possible) is assigned to such events (using an empirically defined threshold).

This post-processing step of calculating the standard deviation of the colour layout bins provided a better user experience. A small test was carried out to investigate the usefulness of this post-processing set. Beforehand while using the *"Previous 3 Days"* approach an average Likert daily rating of 2.486 was obtained. However, after this post-processing step the Likert score for the *"Previous 3 Days"* approach was 3.108). There were fewer irritating events of "SenseCam behind my shirt" scoring highly.

5.1.2 Keyframe Selection for Visual Display

So far we have described techniques to calculate the importance of events. Now the focus is shifted towards discussing methods to select a representative image from each event. We now discuss traditional methods used in the video domain to select a keyframe image, but also discuss our image "quality" approach towards addressing this task.

While there are several sophisticated approaches to keyframe selection within video retrieval, Smeaton and Browne [137] note that the simple approach of taking the middle frame is often favoured as the keyframe image [118, 51] due to both its computational efficiency and good performance. As such, we use this approach as the baseline against which to compare our other approaches. Cooper & Foote [41] have investigated two other keyframe selection methods. The first selects the individual image that is visually closest to all the other images in the event, while the second selects the image closest to all the other images in its event, but which is most distinct from all of the images in other events (these approaches as called "within" and "cross"-event for the remainder of this chapter). These approaches can be computationally expensive though Kehoe & Smeaton [77] have explored taking advantage of graphic processor units (GPUs) to quickly select keyframes using the first method of Cooper & Foote [41].

To determine the optimal keyframe selection method a quality metric is also investigated, as proposed by co-authors in work during our studies [46]. Given that a visual lifelog can contain images of variable quality, it is likely that image quality may play an important role in the selection of an appropriate keyframe. In this work we calculated the quality of each image by combining a contrast measure and a saliency measure (see [46] for further details). Image contrast is a measure of the ratio of the intensity of the brightest colour (white) to that of the darkest colour within an image. A very low or very high contrast value indicates poor image quality, a median contrast measure is preferred. The saliency measure is intended to correlate with "busyness" within an image, and although not exactly an image quality measure, it can be helpful in determining which images have very few salient regions within them which would not be desired for the selection of a keyframe.

In all to select a visual keyframe for all the events, the following six approaches were investigated:

- **Middle Image (Baseline)** - Select middle image;
- **Within Event** - Select the image within the event that is closest to the *average* value of all images in the event;
- **Cross Event** - Select the image within the event that is closest to the *average* value of all the images in this event, but most different to the *average* value of all the images in the other events of that same day;
- **Image Quality** - Select the image with the highest quality;
- **Within Event and Image Quality Fusion** - Select image that is most representative of the event, but which also has a good quality score;
- **Cross Event and Image Quality Fusion** - Select the image that is most representative of the event, which has a good image quality, and is also most distinguishable from the keyframes in the other events;

5.2 Experimental setup

Two distinct sets of experiments were carried out in work relating to this chapter. The first set of experiments investigated the optimal approach to rank events in a day by how

User	Total Num Images	Num Events	Num Days	Daily Duration
1	79,595	1071	35	14h 54m
2	54,823	661	28	12h 29m
3	42,557	405	20	12h 02m

Table 5.1: A summary of the data gathered to be used in event importance experiments.

important they are. The second set of experiments investigated the various approaches to selecting a keyframe image from an event. The composition of these experiments will now be explained in this section.

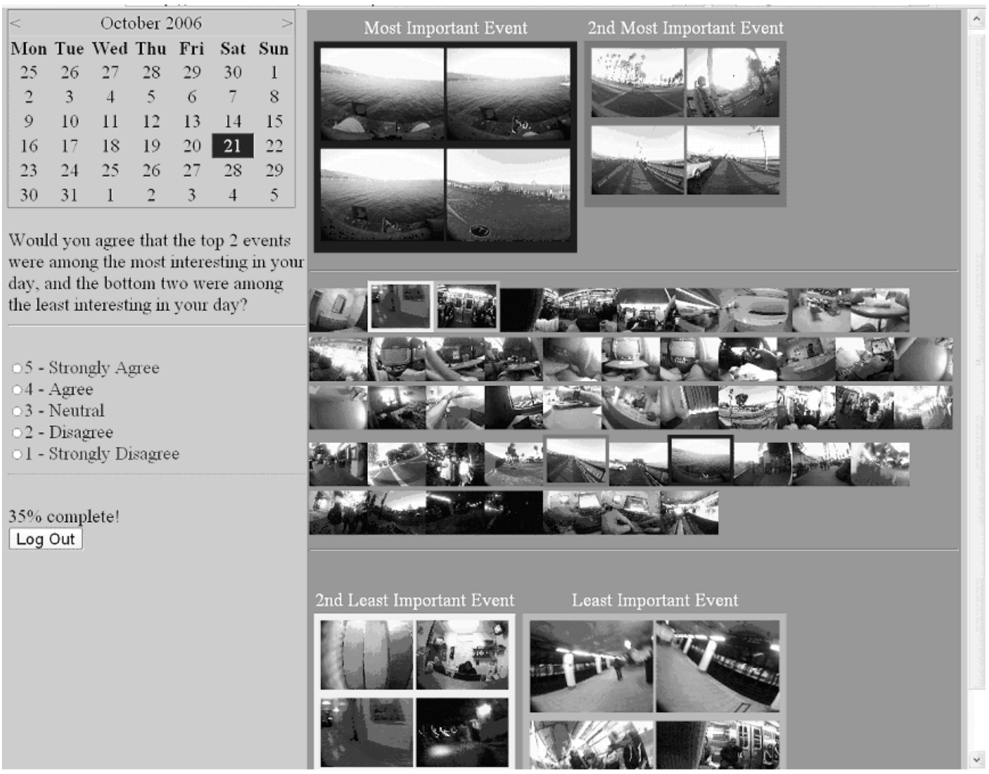
5.2.1 Event Importance Experiments

We now describe the setup of our experiments where six users (information retrieval scientists in their early twenties to mid thirties) each wore a SenseCam over a one month period, collecting 288,479 images, representing 3,445 events in total. However, as novelty calculation is dependent on considering data from previous days to a given day, the experimental dataset was limited to periods of time when a user was wearing the camera for a number of consecutive days. After initial observations, it was decided to only consider periods where a user had been collecting data for at least 14 consecutive days without break. This resulted in a dataset of 176,975 images collected from 3 users, which still represents a large dataset of lifelog images. Table 5.1 provides a detailed breakdown of the 176,975 images captured by the 3 users.

To detect conversational activities (which may indicate important events), we use automatic face detection on SenseCam images. To determine the optimal face detection parameters, the performance of various models was measured against a groundtruth of 1,758 SenseCam images from a day of data belonging to user number 2 in Table 5.1. This data was collected on the 5^{th} September 2006 when the user was attending a conference, so included 289 positive sample images of the user talking to someone else. An average day of SenseCam images can be processed for the presence of faces in approximately 5 minutes on a Dual-Core 2.4GHz machine with 2GB RAM.

Given the unique and personal nature of lifelogging images, the owners of the various SenseCam data collections were the sole judges of the various experiments on their own datasets.

As there is a subjective nature of rating how important an event is in relation to other events, it is very difficult to rank the importance of all events within a day, and to do this for each and every day would present a large annotation burden on users. Therefore given that it is of much interest to determine the *most* interesting events in a given day, in addition to determining the *most* mundane/routine events from a day, a decision was made to present the two most important and two least important (as determined by the approach under investigation) events to the user. Users were then asked to give a Likert judgement on how much they agree with the proposed events as a summarisation of that day.



< October 2006 >

Mon	Tue	Wed	Thu	Fri	Sat	Sun
25	26	27	28	29	30	1
2	3	4	5	6	7	8
9	10	11	12	13	14	15
16	17	18	19	20	21	22
23	24	25	26	27	28	29
30	31	1	2	3	4	5

Would you agree that the top 2 events were among the most interesting in your day, and the bottom two were among the least interesting in your day?

○5 - Strongly Agree
○4 - Agree
○3 - Neutral
○2 - Disagree
○1 - Strongly Disagree

35% complete!

Log Out

Figure 5.2: A screenshot of the application we used to judge the performance of the various event importance techniques.

Figure 5.2 illustrates the application built for users to judge the results from the various

User	Days	Images	Events	Judgements Made	Duplicates Removed
1	35	25,243	360	1,323	837
2	44	67,829	791	2,761	1,985
3	21	42,700	409	1,409	1,045
4	25	40,687	492	1,681	1,271
5	9	18,485	183	639	459
Total	**134**	**194,857**	**2,232**	**7,813**	**5,597**

Table 5.2: An overview of the data gathered for our keyframe selection experiments.

novelty detection approaches on their own data. The top two boxes display a selection of images from the two most important events, while the two blue boxes at the bottom of the screen display a selection of the images from the two least important events in the day (as rated by our proposed system). All the images in the middle are “middle keyframe” images from each event in that day. The two most and least important events are colour-coded among all those event keyframes, and thus their temporal position relative to all the other events in the day is made obvious.

In total 664 judgments on event importance were made by our 3 users and the results will be analysed in the next section.

5.2.2 Keyframe Selection Experiments

Now the focus is shifted from the event importance experiments towards discussing how we set up experiments to judge various approaches to select a representative image from each event.

After selecting six possible approaches to the selection of representative keyframes for events within visual lifelogs, a determination of the effectiveness of each method was required. In order to ascertain this, keyframes were selected for lifelog events using the six approaches and judged by a number of participants. The details of the experimental evaluation are outlined below.

Five individuals (in their early twenties to mid thirties) participated in our keyframe selection experiments and a subset of each participant’s collection was extracted for use within this experiment. The subset represents a continuous lifelog recording of a time period ranging from over one week to a month and a half of the owner's life. 134 days worth of visual lifelog data were used within the experiments, equating to almost two hundred thousand images (see Table 5.2).

For each image within the experimental collection, image quality measures were extracted automatically by combining contrast and saliency as described earlier. The collection was then segmented into 2,232 discrete events using our segmentation method described earlier in Chapter 3 [48]. Potential keyframes for each event were then selected. The middle image from the event was selected as a baseline for comparison within the experiment and a single frame was then selected for each of the following approaches: *Middle Image*, *Within Event*; *Cross Event*; *Image Quality*; *Within Event and Image Quality Fusion*; and *Cross Event and Image Quality Fusion*. This provided up to six potential keyframes per event.

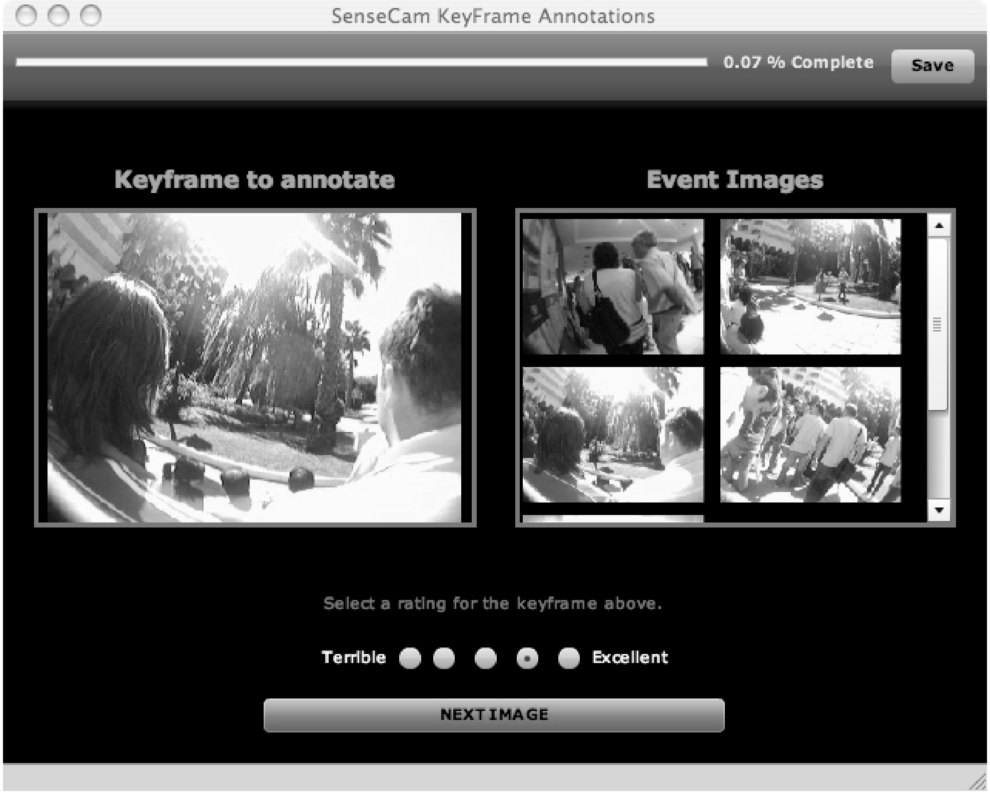
The owners of the original collections were then asked to judge the resulting potential keyframes, rating their suitability as representative frames for the event on a five-point Likert scale. Each collection owner only judged the events and frames they had originally generated. As the same frame will offer the same overview of the concepts contained within an event it was possible to significantly reduce the number of judgments required by each participant. In the case of the same frame selected by more than one approach, the participants only judged that frame once. This resulted in a reduction by 5,597 judgments (see Table 5.2) and ensured consistency in the judgments i.e. the same frame could not be rated differently as the single judgment applies to all approaches.

In order to facilitate the keyframe judgment process, a custom tool was developed (see Figure 5.3). The tool was installed on each participant's desktop computer. Each user completed the judgments on their computer at their leisure. The application provided feedback to the user as to their current progress through the task. At launch or when a judgment is completed, the application selected at random a keyframe to annotate from the pool of remaining un-judged frames. Between judgments a "loading" message was displayed for 2.5

seconds. The random selection and presentation delay were introduced to mitigate against priming and interaction effects.

While making a judgment, the keyframe under scrutiny is presented on the left hand side of the screen while on the right a set of images from the entire event were presented in order to aid the users' judgment. Every 8th image in the event was presented to provide an adequate summary of the event and as a comparative set by which to judge the keyframe.

Users were provided with a set of radio buttons below the keyframe and event images. Users rated the keyframe by clicking on one of these buttons with the mouse or by pressing the corresponding numeric key (1-5) on the keyboard. Once satisfied with the judgment the user pressed the "Next Image" button. This button was only enabled once a judgment had been provided. Users were not allowed to return to a previous judgment.



Select a rating for the keyframe above.

Terrible ○ ○ ○ ◉ ○ Excellent

Figure 5.3: The annotation tool we used in our keyframe selection experiments.

5.3 Results

To recap, two distinct sets of experiments were carried out in this chapter. The results from the first set of experiments investigate the optimal approach to ranking events in a day by how important they are. The results from the second set of experiments investigate the optimal approach to select a keyframe image from an event. These results will now be detailed in this section.

5.3.1 Importance Results

The optimal face detection approach and thereafter the best novelty detection approach will be determined. Finally three systems are compared: 1) The best face/conversation detection approach, 2) The best novelty approach, and 3) Face detection and novelty combined.

5.3.1.1 Best Face Detection Method

The prior state of the art determines the importance of events through detecting events that contain much social interaction, through face-to-face conversations. This is automatically realised by detecting the concentration of faces in an event. In terms of selecting the optimum face detection model, as is illustrated in Figure 5.4, there is a trade-off between precision (dashed line) and recall (dotted line), with approaches sorted in descending order in terms of precision.

In terms of calculating the importance of events through determining the number of images with faces present (relative to the duration of that event), many events are likely to have a score of zero, therefore it is not so important to distinguish between less important events (all the zero-scoring events), but to identify those few events that are important. As a result, a greater emphasis should be placed on the accuracy/precision of actual faces detected by the face detection system. While it may be natural to select the approach with the largest F score (continuous thick black line), it was then decided to use an approach that has a very high precision score (by which the values in Figure 5.4 are sorted). The approach

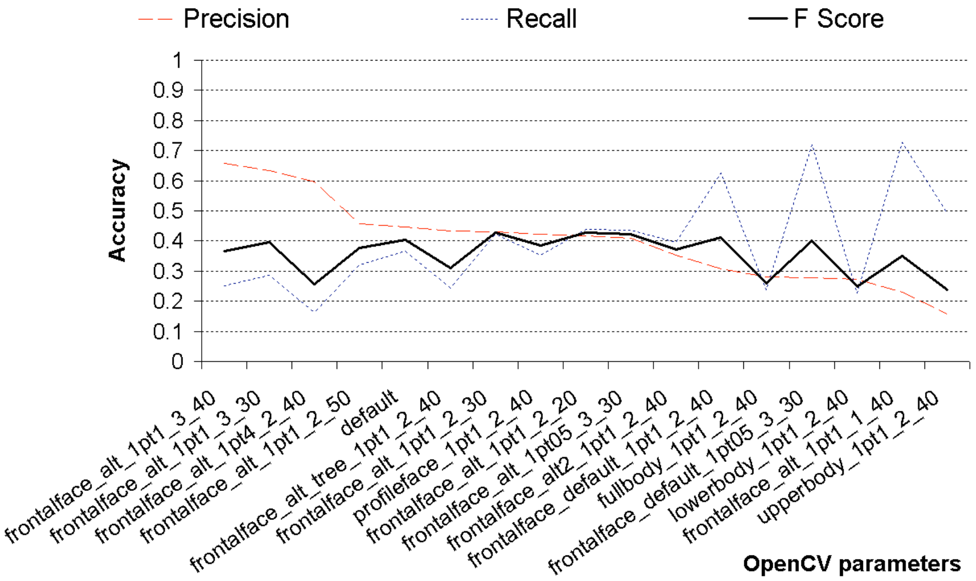


Figure 5.4: A summary of the performance of the various face detection models in the Intel OpenCV library [23].

with the second highest precision score *(haarcascade-frontalface-alt, scaling factor of 1.1, 3 neighbours, and window size of 30 pixels)* is selected as its precision value is very close to the best precision score (0.6336 vs. 0.6577, 3.8% worse), while its recall value is much better than the recall value of the approach with the best precision score (0.2872 vs. 0.2526, 12.0% better).

As discovered in the previous paragraph there are indeed many SenseCam events that have not had the presence of faces detected (763 out of 2137), thus providing further motivation to combine this approach with novelty methods to identify the more mundane events in a given day.

Based on these results it appears that face detection accuracy is *reasonable* with respect to lifelogging, and in particular SenseCam events. As a matter of interest, the in-built passive infrared sensor on the SenseCam device was then investigated to see how accurate it was in detecting faces on this subset. In total 10,528 images were detected as having the presence of faces, using the Intel OpenCV face detection application. Of those images, 8746 also were detected by the passive infrared sensor (83%). Initially this appears very promising however for the 166,447 other processed images the passive infrared sensor was trigged for 135,241 of them (81%). Much of the time the passive infrared sensor can be triggered by the hands of the users in front of the SenseCam device, therefore it can not be considered reliable for face detection.

5.3.1.2 Best Novelty Approach

In the work towards this dissertation a new method of determining the importance of events has been proposed, namely trying to determine how novel each event is. More routine/mundane events (e.g. having breakfast, working on the PC, etc.) have low novelty score, while more visually unique events (e.g. a rock concert) have high novelty scores. To determine the novelty of an event, we compare it to other events that have taken place in the past (so as to build up a model of the individual's routine). Five different "windows" of past events against which the novelty of a given event was calculated, were investigated. Figure 5.5 summarises the performance of these, illustrating the variation in performance for different users for each approach[2]. The middle dots joined up by the thin line indicates the average value across users for each approach. Figure 5.5 addresses 3 questions:

- What is the optimal number of previous days data to consider in the window of novelty detection ?
- Is it beneficial to look only at events that occur at a similar time of day ?
- Is it advantageous to only consider events from the same day of the week ?

Optimal number of previous days to consider ? Taking the overall average of each approach, considering the previous 7 days of data is most beneficial with an overall average Likert score of 3.05 (7 days) vs. 2.94 (14 days) vs. 2.89 (3 days). However, on close examination of Figure 5.5 we recommend that the optimal approach is to consider the previous 14 days of data. The extra processing load to compute this

[2] It is assumed that the users treated the Likert scale as interval variables

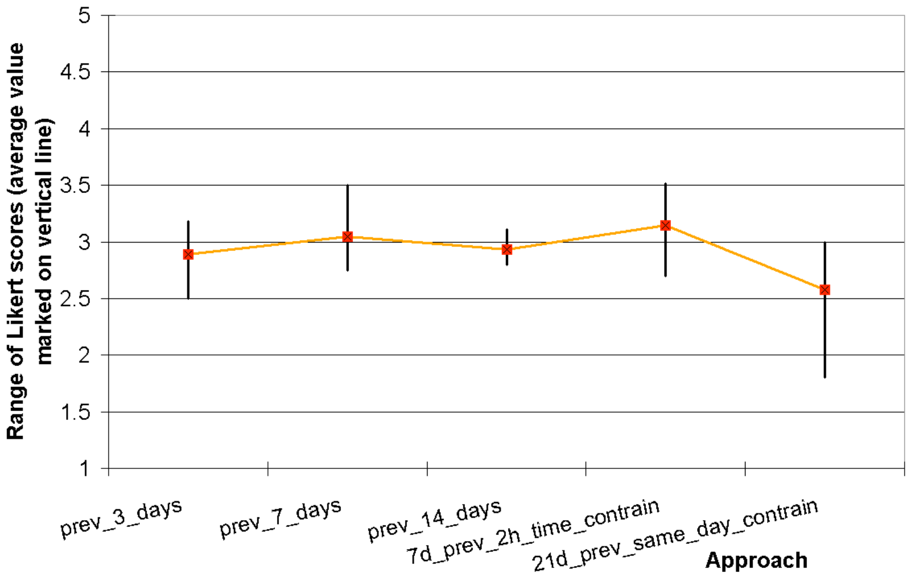


Figure 5.5: A summary of the differences of the novelty approaches in determining event importance.

approach is not a significant drawback (processing for all approaches being practically instant). By considering events from a greater number of previous days (i.e. the previous two weeks) a more accurate model of a user's lifestyle can be considered, while allowing for changes in a user's routine too.

Considering event novelty with respect to time of day ? A comparison is made here between the *7-days-prev-2h-time-constrain* approach which works marginally better than the*prev-7-days* approach in Figure 5.5. An advantage of the time-constrained approach is that it considers the novelty of an event not only based on how different it is to all the other events, but also by considering the time of day when that event took place, e.g. a meeting late in the evening may be very important and unique to the individual, whereas it may not be so important or unique if it occurred during regular working hours. In conclusion it is considered marginally advantageous to calculate the novelty of a given event by only considering comparisons to events from previous days that occur within ±2 hours of the event in question.

Only considering data from same day of the week ? Inspecting Figure 5.5 it is evident that the approach *21-day-prev-same-day-constrain* is the least effective approach. A possible explanation for this is that only 3 days of previous data (e.g. the previous 3 Wednesdays/Mondays/etc.) are considered, and this may mean that an insufficient body of data is accumulated to accurately calculate the novelty of an event. In conclusion it is not advantageous to calculate the novelty of an event against events occurring on the same given week-day only.

In summary, based on the evidence presented above, to consider the novelty of a given event, it should be calculated how dissimilar this event is to all other events that have occurred within ± 2 hours of this event over the previous 14 days. Events are ranked by dissimilarity value, with the most dissimilar being regarded as the most novel/important/unique event of that day.

5.3.1.3 Comparing 3 Proposed Systems

We have introduced a new approach to determining the importance of lifelog events, namely through detecting the visual and sensory novelty of all given events. The prior state of the art addresses this challenge by detecting the concentration of faces present in an event, to indicate conversational events. In this subsection we investigate the benefits of combining both of these approaches together.

As evidenced in Figure 5.6 the *face detection* approach works considerably better on average than the best proposed *novelty* approach. However, even more interesting is the fact that when these approaches are combined together there is an improvement on the current state of the art (face/conversation detection) in this task. The advantage of this approach over the *face detection only* approach is that the *novelty* approach is very good at detecting the routine/mundane events, while the *face detection* is very good at detecting the most interesting events of a given day. When these approaches are combined they complement each other positively. In fact the *face detection* + *novelty* approach performs at least as well

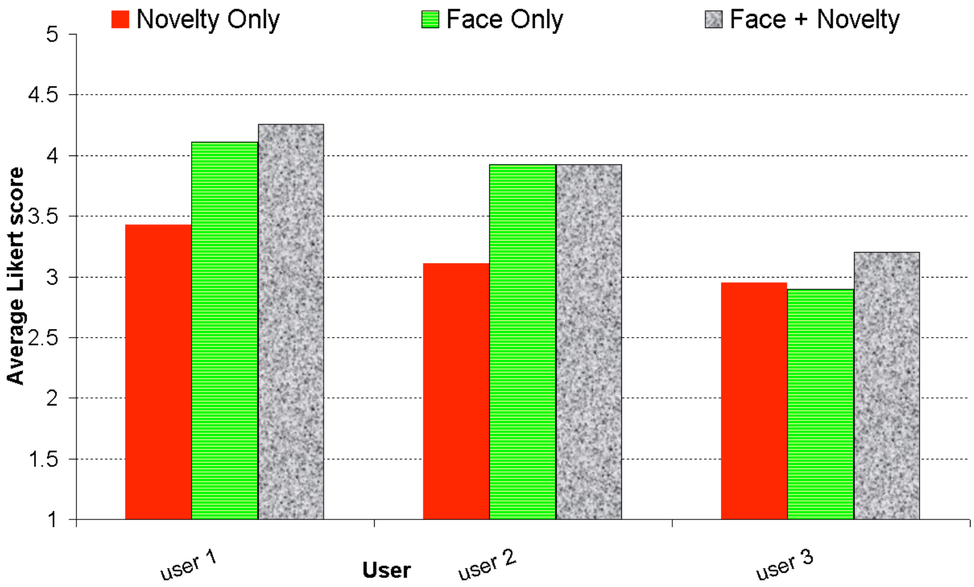


Figure 5.6: A summary of the differences of the 3 proposed final approaches in determining event importance.

as all the other approaches on 66 of the 83 days from all three users, which is almost 80% of the time. Overall, on average, this approach has a Likert score 4% higher (3.89 vs. 3.75) than the state of the art (face detection only).

It is interesting to note that the introduction of the face detection did not have as pronounced a positive impact on user 3 as it had on users 1 and 2. There are 2 factors to explain this. Firstly user 3 was very busy at work abroad during the month he was capturing data, so he did not have as many conversational events during this hectic time. Additionally due to the nature of the light level in certain areas of this user's office, many images were falsely detected as having the presence of a face, as shown in Figure 5.7.

5.3.2 Keyframe Selection Analysis

After determining the importance of an event, which contains an average of almost 100 images, it would not only be useful to emphasise this event to the user, but also to select only one representative/keyframe image from this event so as not to overload the user with visual information. The traditional techniques of selecting a keyframe images include: 1)

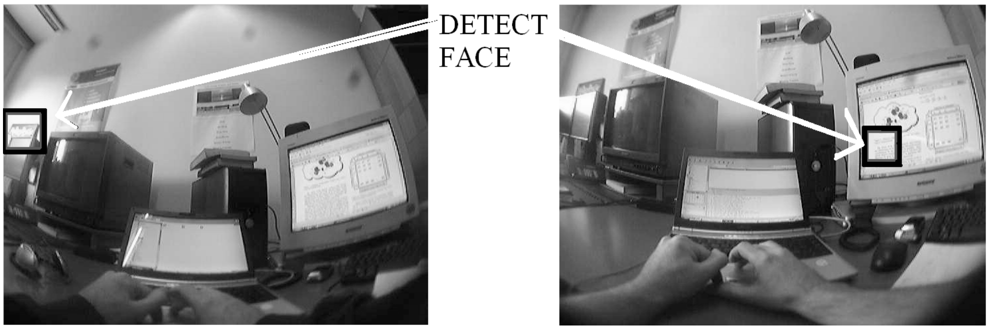


Figure 5.7: An example of an error by the face detection technique on user 3 from table 5.1.

selecting the middle image from the event, 2) selecting the "*within event*" image that is visually most similar to all the images in the event, and 3) selecting the "*cross event*" image that is visually most similar to all the images in the event, but also most distinguishable from all the other images in the other events. In this dissertation we investigate the benefits of selecting the keyframe image through the user of an "*image quality*" metric. The results of these experiments will now be detailed.

Judgements for 13,410 keyframes were provided by 5 users. Ratings provided for each approach were analysed and are presented in Table 5.3[3]. The combination of image quality measures with either "*Within Event*" or "*Cross Event*" selection approaches, prove to be the most effective methods of keyframe selection for visual lifelog collections. Both offer an improvement of 8.4% over the baseline approach. Both approaches offer similar performance but "*Within Event and Image Quality Fusion*" is computationally more efficient and is thus favoured, as it just considers all image in the event rather than all the images from all the events in a given day.

The impact of quality measures within keyframe selection approaches is noteworthy. Selecting keyframes based on the quality features alone outperforms standalone "*within event*" selection and in combination both fused together offer a marked improvement in effectiveness. Quality has a similar effect on the "*Cross Event*" approaches. In combination quality more than doubles the performance gain of "*Within Event*" and "*Cross Event*" over the baseline. This highlights the significance of quality features within visual lifelogs given

[3] It is assumed that the users treated the Likert scale as interval variables

Approach	Avg. Likert Score
Middle Image (baseline)	3.68
Within Event	3.82
Cross Event	3.82
Image Quality	3.91
Within Event and Image Quality Fusion	3.99
Cross Event and Image Quality Fusion	3.99

Table 5.3: A summary of the overall performance of all the keyframe selection techniques.

that they are known to be composed of many poor quality images with high variance in quality over short periods of capture [67]. This variation present in events explains the effectiveness of the quality measures to keyframe selection.

We then further evaluated the approaches taking into account a range of factors including: performance across users' collections, and across the days within the test collections.

Figure 5.8 illustrates the performance of each approach across the users' collections. It is observed that there is a mild variation in the levels to which users have assessed the performance of each approach however they are consistent with the overall findings. In all cases a combination of image quality and either "*Within Event*" or "*Cross Event*" will outperform the baseline. We see that quality was not as effective in the collection of users 2 and 3 but was very effective for users 1 and 5. This would indicate that the quality feature performance is variable when used independently, and is contingent on the user. However, this user effect appears to be tempered by fusing it with other measures.

Figure 5.9 demonstrates that again combination of image quality with either "*Within Event*" or "*Cross Event*" significantly outperforms other approaches for most events. Both approaches prove to be at least as good as the baseline 80% of the time with both offering better performance over 69% of the time. Additionally most approaches perform better than the baseline. It is noteworthy that on initial impressions the quality measure alone as a keyframe selection approach, while performing well, does not appear to provide as consistently good keyframes.

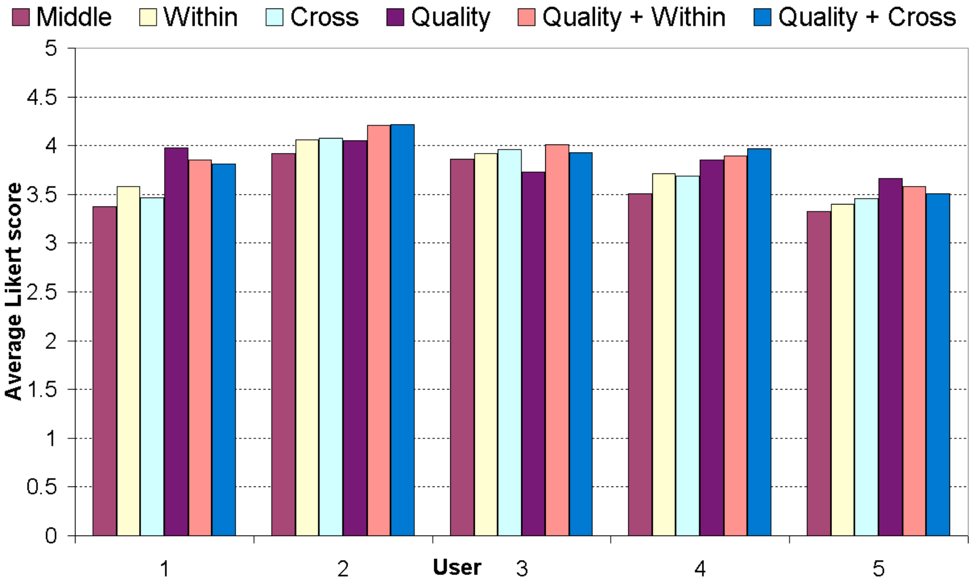


Figure 5.8: A summary of the overall performance of all the keyframe selection techniques for each user.

Lifelog images are typically downloaded and processed on a daily basis. Additionally, a review of lifelog data involves providing a browsable daily summary to the user [29]. As such, consistent effective keyframe selection at the daily level is of importance. To ascertain this, the results were analysed to provide the overall average performance of each approach for each day's worth of events. Detailed exploration of the "Within Event and Image Quality Fusion" approach's performance for each day's worth of events is provided in Figure 5.10, which contrasts this approach with the baseline for each and every day across all users. As can be seen this keyframe approach (by which the graph in Figure 5.10 is sorted) almost always outperforms the baseline middle image.

5.3.2.1 Difficulty in Selecting Correct Keyframe

While the proposed approaches show only a modest improvement over the baseline they are still encouraging. Selecting a representative keyframe for an event can be a difficult task given the challenges presented by a visual lifelog. These include: limitations of event segmentation; events with a high proportion of visual change; or events with high visual

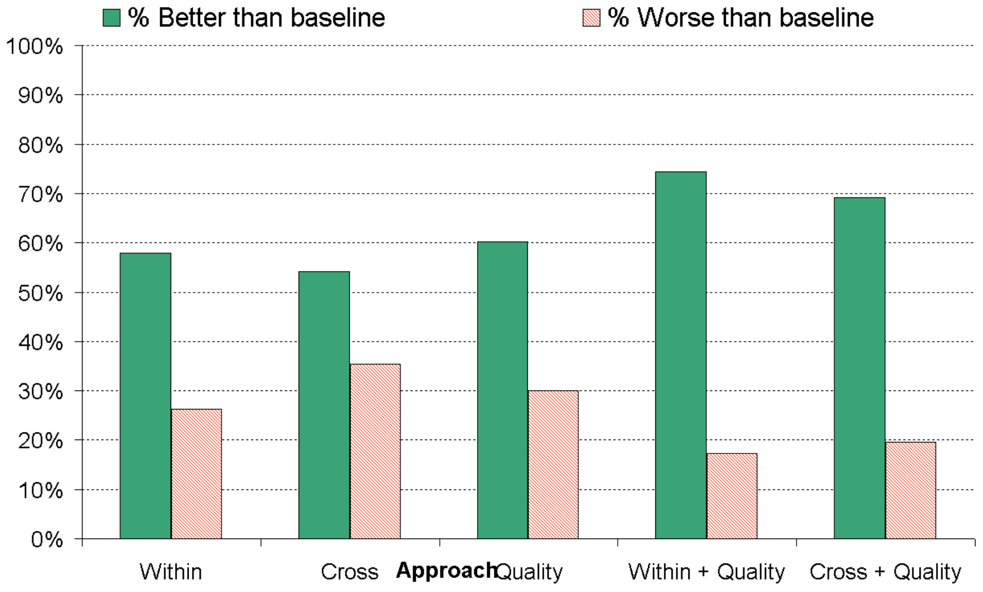


Figure 5.9: An insight into the daily performance improvement offered by each keyframe selection approach over the baseline of selecting the middle image as a keyframe.

change as a result of the nature of the activity. Given the difficulties in selecting keyframes we then decided to explore the performance of the various approaches in events containing a high quantity of visual change.

Issues relating to Event Segmentation

As detailed in Chapter 3, segmenting sequences of lifelog images into distinct events or activities is a considerable challenge. Reasons for this include the fact that wearable devices capture images at a low quality and use a fisheye lens, images are taken at a rate of just 3 images per minute, etc. As a result of these challenges there may be occasions where more than one "activity" is contained within the event, thus making it difficult to reliably select a representative keyframe. Participants of the judgment effort indicated as much in anecdotal feedback following the annotation.

Issues relating to Large Amounts of Visual Change

Depending on the nature of the event a large amount of visual change may be present within the event's frames. As images are captured every 22 seconds (on average) ma-

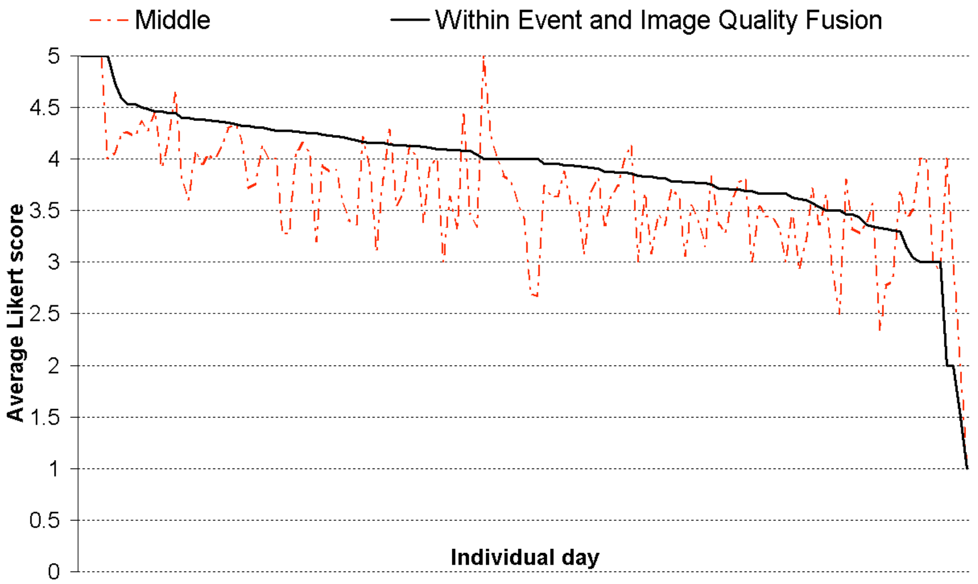


Figure 5.10: A comparison of the daily performances of the “quality+within” vs. “baseline” keyframe selection approaches.

jor changes in the visual landscape can appear from frame to frame within an event. This is true of events which contain motion. For example, an event which represents a wearer walking from their home to the local shop may contain images of walking down the stairs, opening the door, walking down the street, approaching the shop, and arriving at it. With so much change and activity it is much more difficult to select a frame which represents and summarises the salient concepts of the event. Conversely, with some events containing little change and/or motion, e.g. working in front of the computer, it is relatively easy for most approaches to select a representative frame. As such, specific investigation of the performance of selection approaches within these more challenging events should highlight more effective approaches.

5.3.2.2 Selection of Keyframes in Events with High Visual Change

In order to determine the effect of visual change across frames within an event on keyframe selection, the events with large amounts of visual change were automatically extracted for further analysis (based on MPEG-7 features of the images within each event). The colour

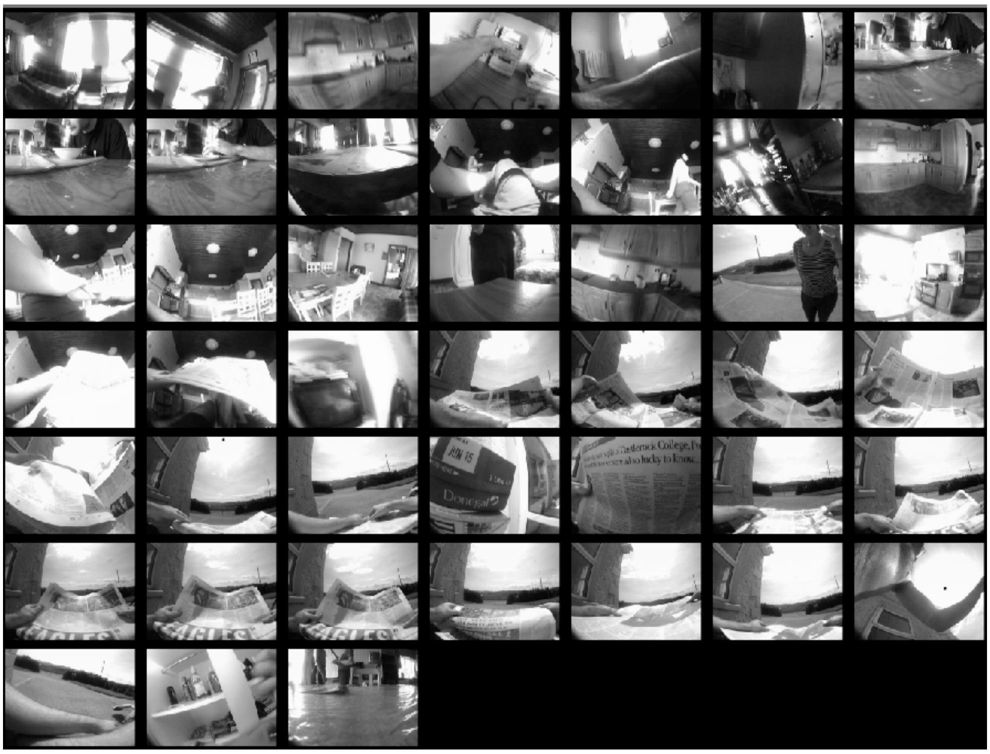


Figure 5.11: An example of an event with a high amount of visual change.

layout descriptor captures the spatial distribution of colour in an image, therefore we calculated the variance of the first bin of this feature for all the images in each event in the test collection. After training on a set of 369 events from two users, a threshold value was empirically defined. Within this training set, it should be noted that the first user only achieved a precision score of 0.48 (66/113) for identifying events with a high degree of variability, whereas the second user's performance yields a much better score of 0.78 (18/23). While not without its limitations, this method does provide a reasonable indication of those events containing a greater amount of visual variation within their images.

The six keyframe selection approaches were then investigated for only those events above the visual change threshold (Table 5.4). It is observed that the scores for events with a high degree of visual variability are lower than the reported performance for *all* events (see Table 5.3). This confirms that there is indeed a significant challenge in selecting keyframes for such events. The quality measures perform well here as these events are likely

Approach	Avg. Score
Middle Image (baseline)	3.31
Within Event	3.43
Cross Event	3.43
Image Quality	3.92
Within Event & Image Quality Fusion	3.73
Cross Event & Image Quality Fusion	3.82

Table 5.4: A summary of the performance of each keyframe selection approach on those events with a large amount of visual changes.

to contain a lot of motion resulting in low quality, blurred or noisy image capture. Again, a combination of quality features and either "*Within Event*" or "*Cross Event*" performs above the baseline and non-fused results.

Performance comparison across collections provided by each user is illustrated in Figure 5.12. While performance of the baseline, "*Within Event*" and "*Cross Event*" operate with reasonable consistency across users, the remaining approaches are subject to a much larger degree of variation in performance. For example, in the case of user 2 all approaches work almost equally well while in the case of user 5 there is a significant difference between image quality and all other approaches. Quality alone appears to be the most effective measure, although, quality in combination with "*Within Event*" or "*Cross Event*" work well when compared against the baseline.

Comparison of the approaches (against the baseline) within each day's worth of (high image variability) events (shown in Figure 5.13) highlights the consistent performance of approaches which incorporate the notion of image quality. Across all the days only taking the events with high inter-image variability into consideration, the approach that improves most on the baseline is the *"Image Quality"* approach. This approach performs better on 60.71% of days, and in fact works *at least* as well as the baseline on 82.14% of days. It is of much interest to compare the results of Figure 5.13 (only events with high inter-image variability) to those of Figure 5.9 (*all* events). The most significant deduction to make is that of all the proposed approaches only the *"Image Quality"* approach performs better (relative to the baseline) on the events with high inter-image variability, than is does

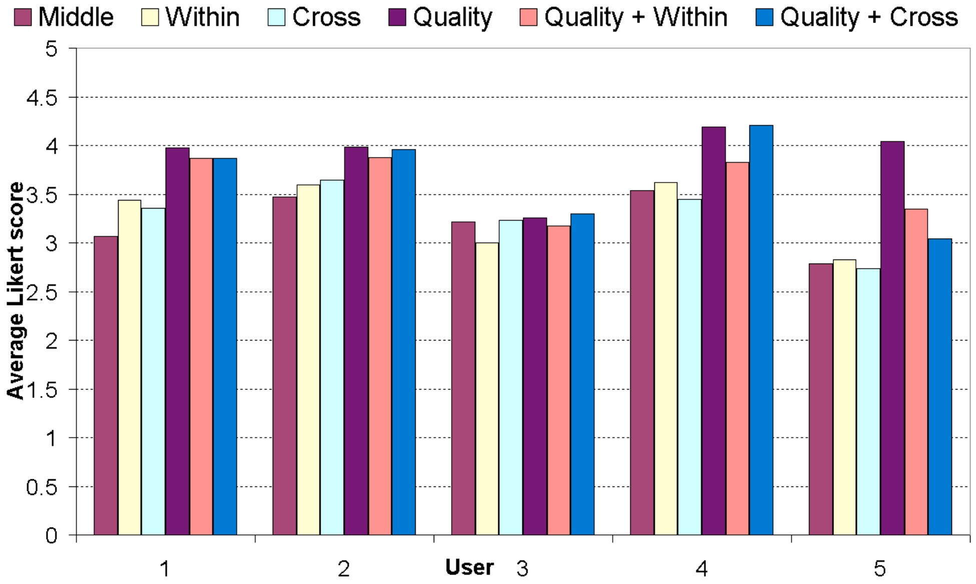


Figure 5.12: A summary of the performance, on each user's dataset, of each keyframe selection approach on those events with a large amount of visual changes.

across *all* events. All the other approaches perform less competitively in those events of great uncertainty, even when fused with the *"Image Quality"* approach. The improved performance gained by using only the *"Image Quality"* approach is further highlighted in Figure 5.14 where it can be seen this keyframe approach (by which the graph in Figure 5.14 is sorted) almost always outperforms the baseline middle image.

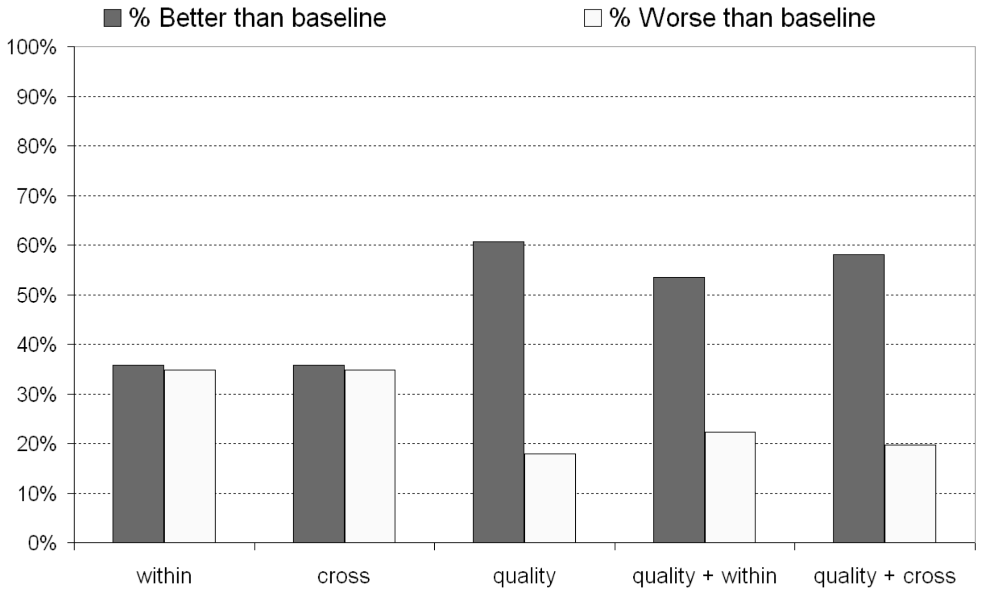


Figure 5.13: An insight into the daily performance improvement offered by each keyframe selection approach, on those events with much visual change, over the baseline of selecting the middle image as a keyframe.

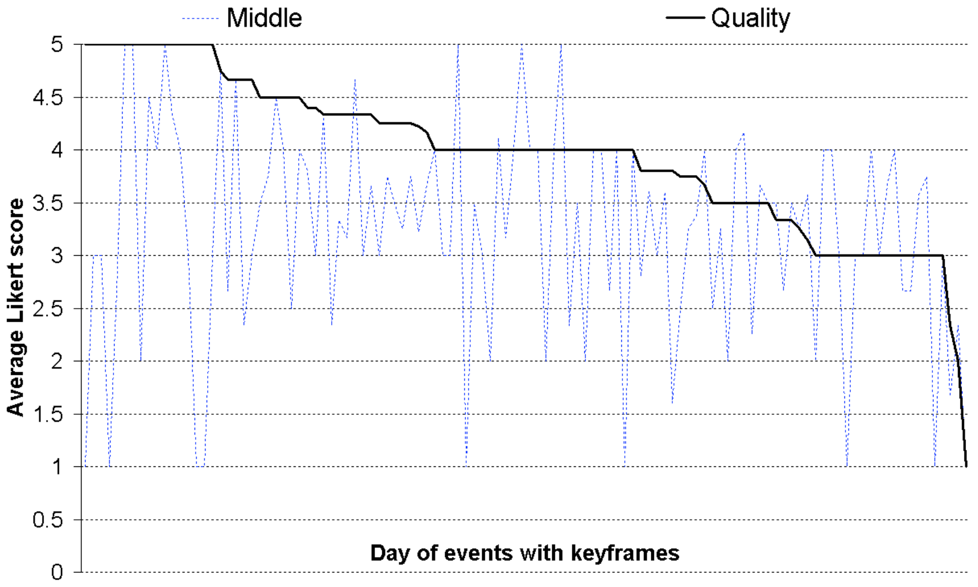


Figure 5.14: A comparison of the daily performances of the "quality" vs. "baseline" keyframe selection approaches, on those events with much visual change.

5.4 Summary

Considering that more distinctive happenings are strongly encoded in the human memory store, in our hypothesis we aim to *"... accurately determine which of those events are most unique ... "* In this chapter we have discussed the usefulness of determining the importance/interest of lifelog events e.g. going to a rock concert is more important/interesting than a normal working event. We have investigated a number of techniques to automatically calculate an importance value for each and every event in our lifelog collection. Also given that events consist of almost 100 images on average, it is useful to provide the user with one representative/summary image of each event, which means they are not overburdened in terms of visual information. We have investigated a number of approaches to automatically determine a good keyframe for any given lifelog event.

A number of approaches to determine the relative importance of lifelogging events from a visual lifelog were investigated on a dataset of 176,975 images from 3 users. The idea of determining the novelty or visual uniqueness of lifelog events has been introduced into this domain for the first time here. The previous state of the art was based on determining events which have face-to-face conversations as the most important events. However, by integrating this with the concept of visual uniqueness or novelty we have developed an approach that works *at least* as well as the previous state of the art 80% of the time and performs 4% better than the previous state of the art overall.

The challenge of selecting a relevant keyframe from a lifelog event is also worthwhile to investigate. Individual events in the lifelogging domain can vary in terms of visual quality, with up to 40% of images being blurred, noisy, light-saturated, very dark, etc. [67]. As such this presents a particular challenge that traditional keyframe techniques struggle to address. We propose a novel technique to select keyframes based on the image within an event that has the highest "quality". One caveat of this approach is that the processing load is greater than selecting the middle image, and in future it is expected to just determine the image quality of a selection of images around the middle of an event. However, this approach does work 6.25% better overall than the baseline approach of selecting the middle image as a keyframe. It was identified that there are a large number of events in the lifelogging domain that can vary in terms of image quality, and just taking those events into consideration the *"Image Quality"* approach performs at least as good as the baseline (*"Middle Image"*) on 82.14% of days, and 15.48% better overall.

5.5 Conclusions and Contributions

- Face detection is good at detecting conversational scenes in one's lifelog, which are in turn good indicators of the more interesting events for an individual to review.
- We have introduced the concept of novelty to the domain of lifelogging. This is effective in identifying routine activities. To calculate the novelty of a given event, the best event comparison techniques from Chapter 4 are recommended. It should be investigated how dissimilar events are to other events around the same time of day from the past fortnight.
- We have discovered that the *Image Quality* technique performs very well on events containing a large amount of visual change. However the baseline approach of simply selecting the middle image still performs quite competitively, while being computationally much less expensive. An interesting area of future work would be to examine the effectiveness of selecting the highest quality image from a selection of images around the middle of events.
- Papers have been published on event importance at CIT 2008 [49] and on keyframe selection at CIVR 2008 [46].

Chapter 6

AUGMENTING SENSECAM EVENTS

> *In order for visual lifelogging technologies to effectively provide autobiographical memory retrieval cues it is necessary to automatically:*
>
> ...
>
> - *augment images from the low-resolution wearable device with higher quality images from external data sources.*
>
> ...

The previous chapters have dealt with scenarios of people continuously wearing SenseCams for entire days and over extended periods of time. However, there will also be a large number of people who may not be comfortable in doing this as Bannon feels that it is important to forget certain moments [12]; and Nack also espouses similar concerns [104]. Still though, there are many people who enjoy taking photos of specific events e.g. a rock concert, a football match, a holiday abroad, etc. Gemmell *et. al.* note that '*... there is a strong demand for capture of life experiences, whether in photos, videos, or written accounts. However, few people want to miss the experience in order to be the camera operator ...*' [61]. An advantage of a passive capture device like the SenseCam is that it requires little photo capture effort on the part of users on their tourist trips, or trips to big events, thus helping them relax and enjoy the experience of being there. Afterwards they can review their pictures to help recreate the experience of *being there*.

Even if we do perfect event segmentation and browsing, we believe that it would also be useful to augment events with images from external sources. Consider an assistive tool for tourists visiting, for example, a certain historical famous building. After reviewing the event later in the evening, a tourist may perhaps be interested in viewing other images of that building taken by other people, which will provide additional memory retrieval cues. Another potential use would be of an individual attending a music concert or a sporting event, and having the ability to view images taken by other people of that same event. The SenseCam images are taken at quite a low quality, however many images uploaded to photo sharing websites by users with modern day digital cameras will be of a much higher quality. Also they may be taken from a different perspective which the SenseCam user may enjoy viewing (and again provides extra memory retrieval cues). These external images augmented to the low resolution SenseCam images can translate to a more enjoyable user experience when reviewing certain SenseCam events.

An obvious application area resulting from this would be an aid for tourists to review their trips, e.g. Aiden wears his SenseCam on a trip to New York and while there he travels to the Statue of Liberty, Times Square, and also to a baseball game. When he arrives back in Dublin he downloads his SenseCam images to his PC and enjoys reviewing images of his trip. However, he can augment his trip to the Statue of Liberty, as not only can he see his own SenseCam images, but he can also view images taken by entire communities of users from sources like Flickr, Panoramio, YouTube, and the MSN & Yahoo! image search engines. Similarly Aiden also really enjoys looking at pictures taken from multiple perspectives at Times Square, which adds to his experience of “re-living” his trip to New York. Finally in looking for augmented images of the baseball game, it is important that the images used are of the actual game that Aiden was at, as opposed to other events that take place at the same venue but which were of no personal meaning to Aiden.

As a result, this chapter is focused on realising the goal of pervasive user-generated lifelog content, using visual content passively captured by an individual and then augmenting that with content collected by other individuals. There will be a number of research challenges associated with this which we address in the next section

6.1 Image Data on the Web

Much attention has been focused on the changing dynamics of the World Wide Web, with the term “Web 2.0” now being widely recognised and understood. In the past much of the content on the web was generated by a small number of professional writers or content generators, with the majority of users only being able to view the generated content. However, with the advent of “Web 2.0” technologies, content on the web can now be generated by *everyone*.

An example of a “Web 2.0” multimedia web site is Flickr[1], a web site allowing the upload of images taken by users themselves. As is reflective of many “Web 2.0” sites, the volume of users and photos on Flickr is astounding. Van Zwol reports that Flickr has 8.5 million registered users, and at peak times there are up to 12,000 images served per second; the record number of photos uploaded per day is more than 2 million [150] ! Flickr also provides users with the opportunity to specify the location of their uploaded photographs by dragging them onto a map and thus automatically appending the GPS coordinates. There are now over 65 million ‘geotagged’ images on Flickr (at time of writing) and the growth rate has been phenomenal (approximately 500,000 new geotagged photos uploaded per week as illustrated in Figure 6.1). Users also can easily tag their uploaded images, and given the significant user base on Flickr this can very quickly create a vast collection of tags which are potentially useful in retrieval.

However Flickr is not the only source of user-generated images. Google Earth also allows users to upload images, which are then accessible on the Panoramio website[2]. *All* of

[1] www.flickr.com

[2] www.panoramio.com

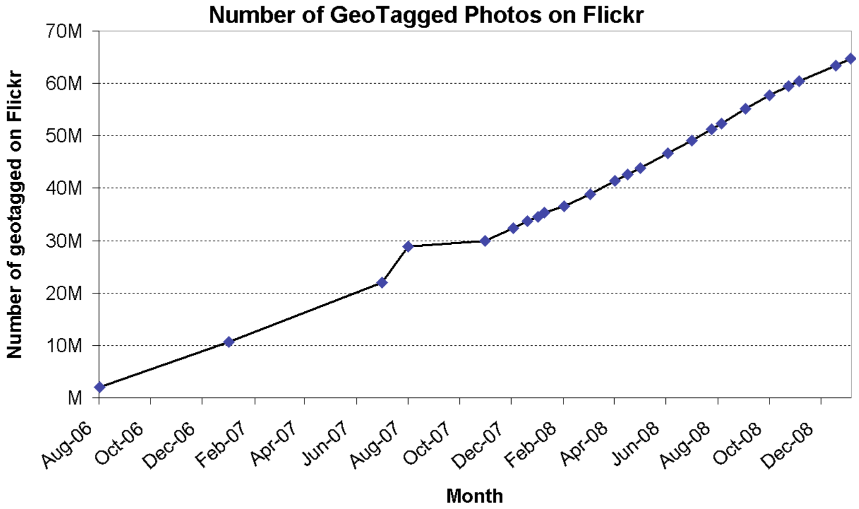


Figure 6.1: An illustration of the number of geotagged photos on the Flickr "Web 2.0" website.

these images are geotagged. Another useful web site is YouTube[3], where millions of users upload *their* own videos. In recognising the shift towards user generated content on the web, Gill *et. al.* report that there are 100 million video views per day on YouTube (which is 60% of the total watched videos on the entire Internet), and 65,000 new videos are uploaded per day [62]. Like Flickr, YouTube also allows content to be tagged by users.

While a very large number of images are now uploaded on the web by many users, the more traditional sources still publish images which are searchable by the dominant search engines (i.e. Google, Yahoo![4], and Microsoft[5]). These search engines index billions of images (e.g. in a clustering paper Liu, Rosenbery, & Rowley of Google used a set of 1.5 billion images from their image index [91]). These indices can now be queried by using publicly available API's, as can the "Web 2.0" image/video sites.

Prior work in augmenting tourist photographs with other material has been carried out against small-scale databases in which much image processing is required. Blighe *et. al.* intelligently identify museum images on the fly, but against a small fixed database of images

[3] www.youtube.com

[4] search.yahoo.com

[5] www.msnsearch.com

[21]; Chevallet, Lim, & Leong identify 101 Singapore tourist locations, but on a dataset of just 5,278 images [36]; while O'Hare *et. al.* have completed similar work and again on a fixed dataset [115]. Since we expect lifelogging devices to become more commonplace in future, and given the phenomenal growth of multimedia content on websites, we introduce the idea of augmenting lifelog events with "Web 2.0" content. We investigate aggregating these small contributions over an enormous scale of users, which can thus automatically enrich the experience of individuals reviewing their trips or events, by providing them with a large number of relevant items of information mined from millions of other individuals.

To realise the goal of augmenting lifelog events with images from external sources of information there are a number of challenges to be addressed. In order to augment a lifelog event with external images we need to know where and when it took place. Thus it is necessary to record location (via handheld GPS device) and time information along with the captured SenseCam images. Once this is done then the user is given the option of whether to constrain augmented images to a certain time or to allow images taken at any time to be used for augmentation. The time associated with images taken of an historical building is not so important, whereas the time when images are taken at the location of a football match will be very important as the user will want images captured by other people at that particular match, not of other matches in the same stadium! GPS information can be captured to record precise details of the user's location. With the knowledge of where an event is taking place it will then be possible to query an external data source for pictures only within an arbitrary distance of the location of where the user was for the particular event.

However in the case of an historical building for example, there may be many close-up pictures taken by other people of their friends at that location. We have investigated, at query time, carrying out a form of image similarity matching between the SenseCam event keyframe image, and each retrieved image from the external data source. If a retrieved image is sufficiently similar in a visual sense to the SenseCam image, based on a certain threshold value, then it will be displayed to the user as a potentially relevant image to augment the lifelog event. This could result in the user being returned with a larger number of highly ranked relevant images from external data sources.

6.2 Augmenting Lifelog Events with Web Images

In this section we will provide details of our techniques which realise the goal of automatically augmenting lifelog events with images taken by many other users (and placed on the Internet).

Figure 6.2 provides an overview of our approach. The process is broken up into 3 stages: 1) retrieve all geotagged images from the same location, with the option to also constrain by time; 2) inspect the tags associated with those retrieved images to construct a new text-only query; and 3) using the text query, retrieve images from several APIs on the web. We will now provide a walkthrough example before discussing how we address each challenge.

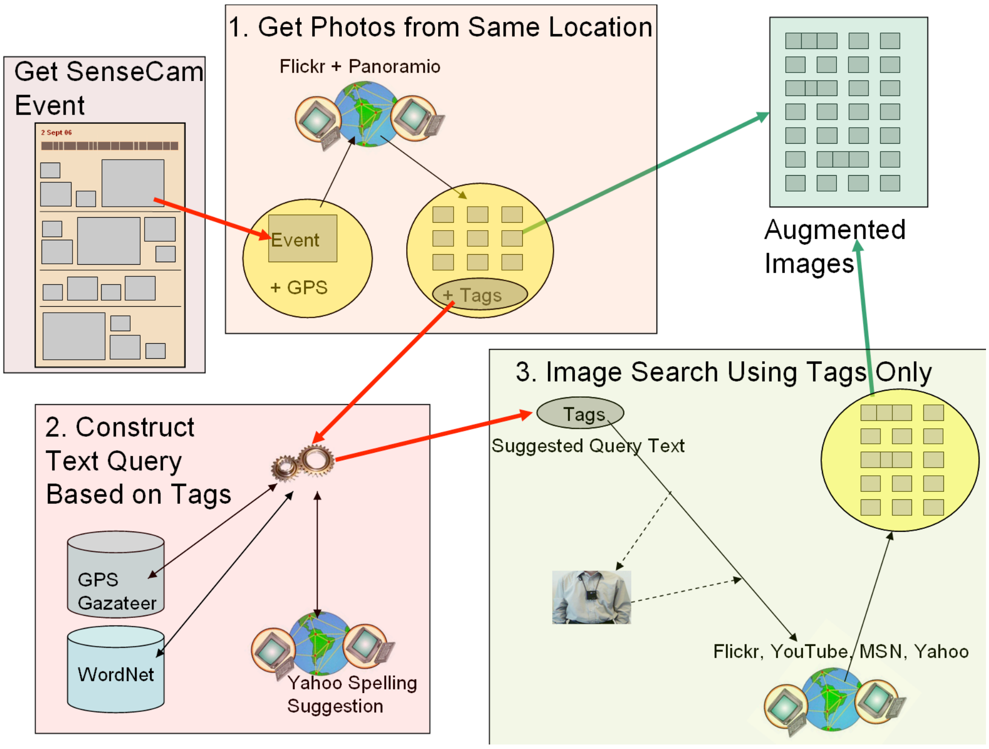

Get SenseCam Event
2 Sept 06
1. Get Photos from Same Location
Flickr + Panoramio
Event
+ GPS
+ Tags
Augmented Images
2. Construct Text Query Based on Tags
GPS Gazateer
WordNet
Yahoo Spelling Suggestion
3. Image Search Using Tags Only
Tags
Suggested Query Text
Flickr, YouTube, MSN, Yahoo

Figure 6.2: An overview of our lifelog event augmentation processing.

6.2.1 Walkthrough of Processing Stages

Before detailing how we realise the processing steps of Figure 6.2 we will now walk through a high level overview of a sample event that one of our users wanted to augment, so as to add to his experience of reviewing his trip in Singapore.

- **Step 1: Get Photos from Same Location** Consider an event where one of our users was in Sentosa Island in Singapore (see sample image in Figure 6.3). To augment this event we firstly retrieve a sample of up to 100 images taken in the same location which were made available on the Flickr and Panoramio websites (see 20 random images sampled in Figure 6.4). In this instance we take 50 sample images from each geotagged source.

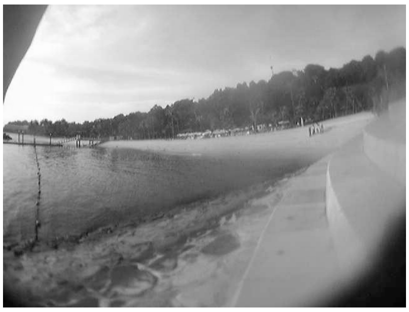


Figure 6.3: A sample SenseCam image from an event at Sentosa Island, Singapore.

- **Step 2: Construct Text Query Based on Tags** Knowing the GPS co-ordinates where the event took place, we can automatically extract the region and country name using a gazetteer [115], which gives us “Pulau Palawan — Singapore (general) — SINGAPORE”. We then make use of WordNet [102] to expand the list of placename terms to include “Republic of Singapore” and also “Singapore Island”. Knowing the region and country names, we ignore such data in the user tags as this does not add any more useful information. Therefore we only consider the tags that *aren’t* of any variations of the region or country names, and count how many unique individuals used the remaining tags (see Table 6.1).

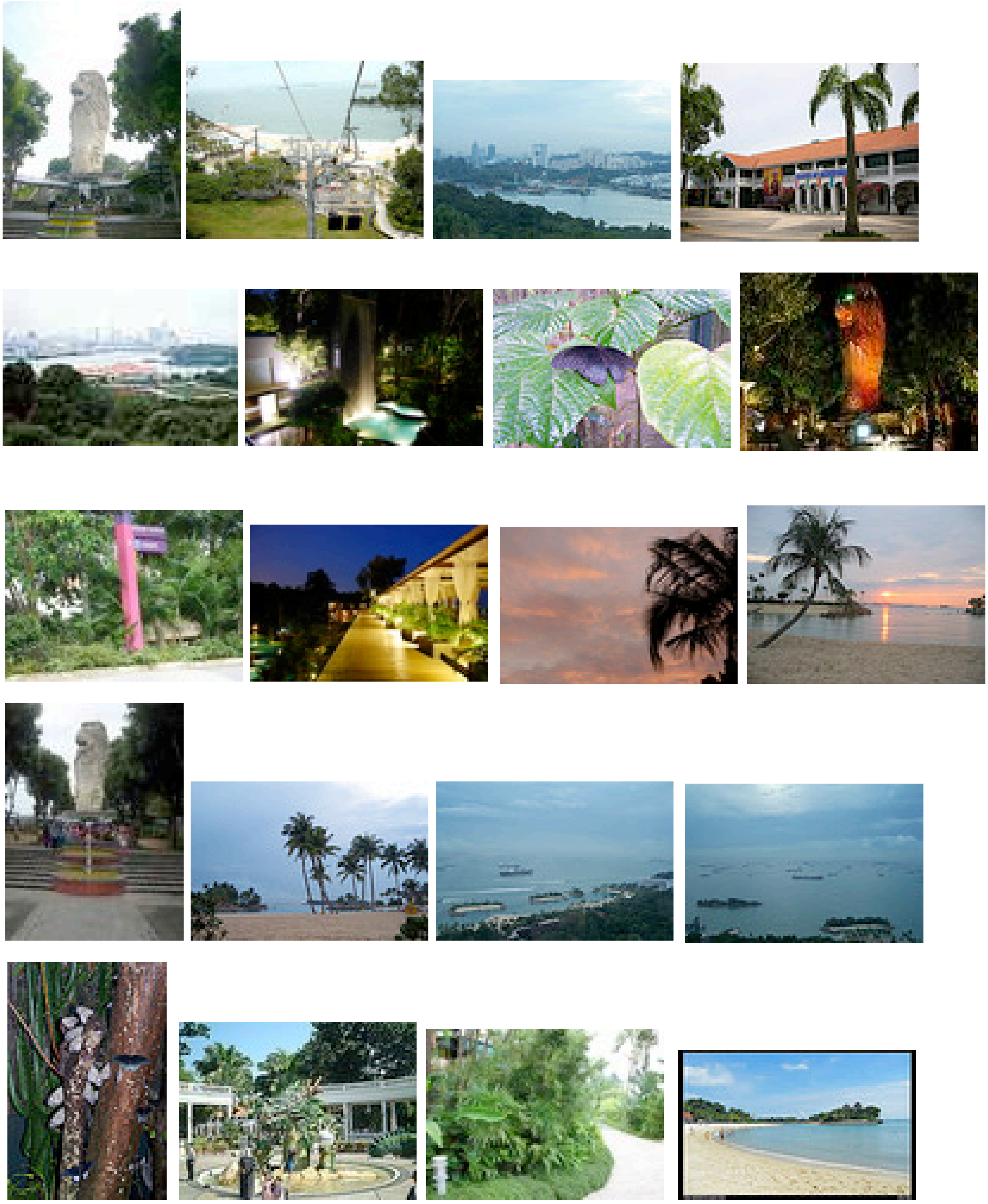


Figure 6.4: Sample geotagged images of Sentosa Island, Singapore from the Panoramio and Flickr websites.

Tag	Num Users	Tag	Num Users
Sentosa	27	from	3
Island	10	of	3
Merlion	10	Island,	3
Beach	5	Garden	2
at	4	flowers	2
geotagged	4	Dragon	2
The	3	Butterfly	2

Table 6.1: Sample tags associated with geotagged images returned for the "Sentosa Island" query event.

The 4 most commonly occurring tags are selected and also the region name (by knowing the GPS co-ordinates), and in the "Sentosa Island" example, we are left with the terms: "Sentosa", "Island", "Merlion", "Beach", "Singapore". Finally we spell check each of these terms using Yahoo! spell checker, so as to allow us retrieve the maximum number of potentially relevant results when performing the text-only search.

- **Step 3: Image Search Using Tags Only** Finally we search the MSN, Flickr, Yahoo!, and YouTube sites with the text query of "sentosa island merlion beach Singapore". For this event 39 of the geotagged images (from Flickr and Panoramio) were relevant, but through the automatically generated textual query, we have added another 43 new relevant items (from Flickr, MSN, Yahoo!, and YouTube) to embellish our user's experience of reminiscing on his time in Singapore.

Having walked through an example, we will now go through each of these processing stages (illustrated in Figure 6.2) in more detail.

6.2.2 Download Photos from the Same Location

With the SenseCam device which our users use for lifelogging, the time of capture of images and environmental sensor readings are automatically recorded. With mobile devices now increasingly carrying GPS loggers (e.g. Nokia N95[6]) it is possible to automatically record the GPS location of our whereabouts on the device itself. Using logs from both of these

[6] http://www.nseries.com/products/n95/

devices, SenseCam and mobile phone, it is possible to query the Flickr and Panoramio (i.e. Google Earth) websites using location information, through publicly available APIs. This returns a set of Flickr & Panoramio photos taken at the same location as the images from the SenseCam event in question.

For some of our lifelog events, the time of the event may be critical, e.g. attending a sporting event or rock concert, whereas for other events, such as visiting the Statue of Liberty, it is not. On inspection of a sample set of 12 major time-critical events, such as sporting events, we noted that 53% of photo uploads of those events occurred within 2 days of the actual event taking place (note some examples of this on the right hand side of Figure 6.5). We constrain images for such events to be within ±5 days of our lifelog event. In our work we ask users to specify whether an event should be time-constrained or not e.g. FA Cup final in Wembley is time-constrained, while photos of Big Ben generally are not. However, in future this process can be automated as Rattenbury, Good, & Naaman have developed an approach that can automatically identify if an event is time critical or not [121].

In selecting the set of query tags to use in the next processing stage it is necessary to initially gather some seed images which will have associated tags. We investigated 3 possible approaches in selecting seed images, namely:

1. **T100:** Forwarding tags from the top 100 geotagged returned results (as ranked by Flickr and Panoramio[7])

2. **T25:** Forwarding the tags from the top 25 results which may be quicker to process

3. **M25:** Using the tags from the top 25 geotagged results that are visually most similar to the lifelog keyframe image (using the set of 100 top-ranked geotagged images returned by Flickr and Panoramio). This requires us to extract image descriptors for 100 Flickr/Panoramio images on the fly, which has a computational overhead. We determine visual similarity between images using the scalable colour MPEG-7

[7] Up to 50 results from Panoramio API (when possible) and the others (50+) from Flickr API

image descriptor. In work carried out in the previous chapters of this thesis we used 4 MPEG-7 image descriptors (*colour layout, colour structure, scalable colour, and edge histogram*), but as it is intended for image descriptors to be extracted on the fly, we select the scalable colour descriptor because it performed well in retrieval experiments, and also because it can be computed quickly.

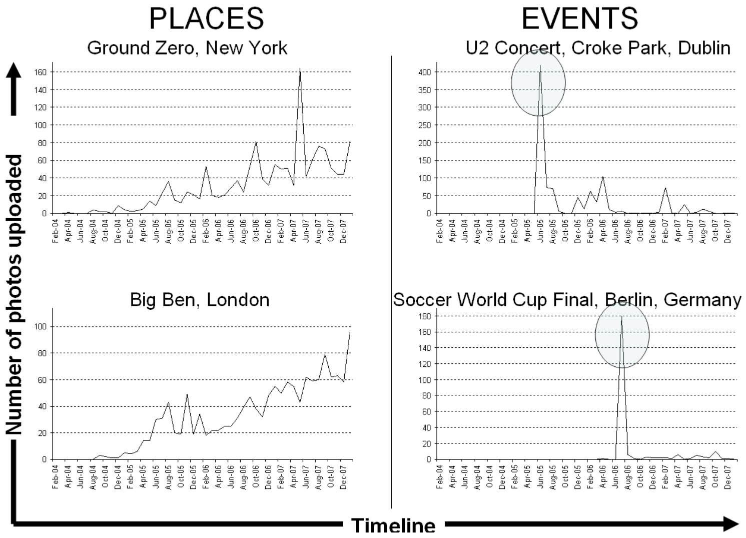


Figure 6.5: An insight into the differing characteristics of location driven images vs. location+event driven images.

6.2.3 Construct Text Queries Based on Tags

In some circumstances it may be acceptable to return only the Flickr and Panoramio geotagged images to augment an event. However, not all images are geo-tagged. In taking a set of 26 sample queries of famous locations (e.g. Eiffel Tower) and events (e.g. FA Cup Final) we discovered that only 22% of the photos of these *landmarks* are geotagged (see Table 6.2). In fact, given that there are over 3 billion images in total on the Flickr site[8], we can infer that only 2% of *all* photos on the entire Flickr database are geotagged. Therefore in wanting to return as many relevant images as possible for event augmentation, we must also search other sources not only for geotagged items, but for *all* items which are relevant.

[8] http://blog.flickr.net/en/2008/11/03/3-billion/

Query	No. Unique Users	No. Geo Images	Non-Geo Images	Total Images
Sugar Loaf, Rio De Janeiro	841	1,032	2,961	3,993
Niagara Falls Canada	480	237	3,263	3,500
Victoria Falls Livingstone Zambia	172	388	2,080	2,468
Sistine Chapel the Vatican	790	534	1,641	2,175
Big Ben London	1,120	323	1,175	1,498
Opera House Sydney Australia	641	540	956	1,496
Eiffel Tower Paris	771	456	1,038	1,494
Ground Zero New York	613	205	1,283	1,488
Pyramid Giza Egypt	358	323	677	1,000
Golden Gate Bridge San Francisco	342	440	560	1,000
U2 Croke Park Dublin	64	41	931	972
Namdaemun Gate Seoul	230	255	619	874
Monaco Grand Prix 2007	68	173	420	593
Taj Mahal Agra India	137	171	329	500
Tower Bridge London United Kingdom	162	191	309	500
Petronas Towers Kuala Lumpur	227	159	341	500
Champions League Final 2007 Athens Milan Liverpool	55	34	340	374
Japanese Grand Prix 2006 Suzuka	9	185	173	358
Fa Cup Final Wembley 2007	36	48	307	355
Super Bowl 2007 Dolphin Stadium	34	5	305	310
Soccer World Cup Final Berlin Italy France	67	8	219	227
Real Madrid Barcelona Camp Nou	13	8	180	188
AFL Grand Final 2007 Geelong	16	0	143	143
African Nations Cup Final 2008	13	0	72	72
Heineken Cup Final 2006 Cardiff Munster	7	12	45	57
Copa America Final 2007 Venezuela Brasil Argentina	4	1	35	36
overall	**7,270**	**5,769** **(22%)**	**20,402** **(78%)**	**26,171**

Table 6.2: This table provides an insight into the percentage of geotagged images, from random sample of 26 queries.

Since we are initially able to search for geotagged photos taken at the same location as the user's lifelog event (as described in section 6.2.2), it is then possible to inspect the tags associated with those images, and intelligently construct a new text query. Textual annotations are powerful descriptors of image content, and Berendt & Hanser claim that tags are not metadata, but are for all intents and purposes an extension of the content itself [16]. In using tags we are not restricting the augmented image set to just those that are geotagged. However, as Schmitz notes, user generated tags often contain bad spelling or no spaces between words [129]. Given all the unique tags from each Flickr/Panoramio user, we "clean up" those tags in two steps [9]:

- Firstly many photo tags in Flickr and Panoramio are of the actual country name or region name, which is not so useful, as we already have GPS information recorded. Also as the majority of tags contain country/region names which create a lot of noise. Therefore we want to ignore occurrences of those country and region name tags. Knowing the GPS location where the event took place, we use a gazetteer [115] to automatically extract the associated country and region names e.g. "New York, United States" and then we use WordNet [102] to expand the list of possible tags that users could have used to mark this place, e.g. "United States" expands to: "US", "USA", "United States of America", etc.

- Of the remaining tags, the 4 most commonly occurring across users are then considered as these are likely to be most relevant to a given event query. Many times those tags may be misspelt or contain spaces, therefore we use the Yahoo! spelling suggestion API to correct any erroneous tags e.g. "statueofliberty" becomes "statue of liberty".

In essence this process is a form of pseudo relevance feedback whereby an initial query is formed through searching by location, and a new query is then constructed by looking at

[9] While initially it may seem that performing the step of getting initial seed geotagged images (as described in section 6.2.2) may be misguided as only 2% of the collection of Flickr images are geotagged; it should be remembered that this 2% represents an enormous collection of over 65 million images!

commonly occurring tags in the initial set of geotagged images.

6.2.4 Image Search Using Tags Only

Given a set of relevant tags, we can now construct text queries, allowing us to search for images from websites with publicly available API's such as YouTube, MSN, and Yahoo!, in addition to the images we get from websites with geotagged material like Flickr and Panoramio. This provides 5 potential sources of images, from which we can augment any lifelog event. We allow our users to alter the suggested textual query in certain instances, and then we provide a number of candidate images from the 5 different sources to supplement *their* experience of reviewing *their* lifelog events. In the remainder of this chapter we explore how these facilities are used in practice.

6.3 Experimental Setup

To evaluate the effectiveness of our proposed event augmentation approach, and also to help answer a number of the research questions we posed, we now describe the setup of our experiments where eleven users collected lifelog data over a period of 2 years. Many wore the SenseCam device sporadically (generally at times of interest, e.g. conference trips, holidays), and in total we collected 1,937,770 images which were segmented into 22,911 events (see Table 6.3).

Users were asked to select a number of events that they would like to have augmented with other images from external sources of data. They were presented with an event based browser (as illustrated in Figure 6.6) to sift through their SenseCam images. The calendar allows the investigator to browse to a day of interest. The vertical column of images in the centre then displays each event for a selected day, and once clicked all images from the event are shown on the right of the screen. Users can select events for augmentation by clicking on the relevant radio button as to whether it is "event specific" or "place specific", and then by clicking on the "tag event" button. A list of tagged events is displayed under

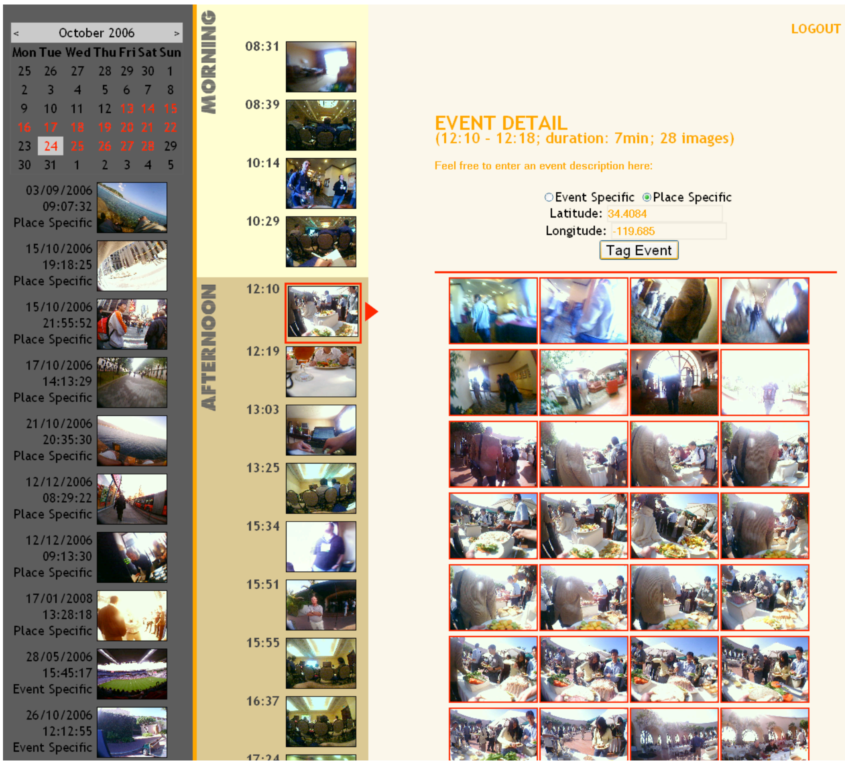
User	Num Days Worn	Total Num Images	Total Num Events	Place Tagged	Time Tagged
1	614	1,686,424	19,995	5	4
2	60	92,837	1,182	8	2
3	21	44,173	443	9	1
4	3	3,437	39	0	1
5	23	40,715	505	5	1
6	24	38,106	465	5	0
7	8	18,485	169	7	0
8	6	8,296	57	3	1
9	1	2,046	20	6	0
10	1	667	8	6	0
11	2	2,584	28	2	1
Total	**763**	**1,937,770**	**22,911**	**56**	**11**

Table 6.3: A summary of the data collected by our users for the event augmentation experiments.

the calendar on the left hand side of the page. In total, 67 events were selected by our 11 users to be augmented (see place and time "Tagged" columns in Table 6.3), with 11 of those events time-specific (e.g. such as a sporting event, rock concert, etc.).

The search for augmented images was then run for each event in our dataset. Two variations of our system were explored; where text from image tags was used as the query; and where the user had the opportunity to amend the suggested query text. Users were presented with a screen full of candidate images, 10 from each data source, presented in random order, for event augmentation (see Figure 6.7) in which they had the option to select images relevant to their lifelog event, which were in turn highlighted with a green background. By moving the cursor over a candidate image the user was provided with a text description taken from the webpage that the image belongs to. Below each augmented image there is a "Link" button which opens the original source of that image in a new browser window. 4,963 candidate augmentation images/videos were retrieved for presentation to users for their judgement.

After the judgements were complete, users were questioned on: the usefulness of the suggested query text on semi-automatic runs; the general effectiveness of the system; and their satisfaction with the augmentation process overall.



< October 2006 >

Mon	Tue	Wed	Thu	Fri	Sat	Sun
25	26	27	28	29	30	1
2	3	4	5	6	7	8
9	10	11	12	13	14	15
16	17	18	19	20	21	22
23	24	25	26	27	28	29
30	31	1	2	3	4	5

03/09/2006 09:07:32 Place Specific

15/10/2006 19:18:25 Place Specific

15/10/2006 21:55:52 Place Specific

17/10/2006 14:13:29 Place Specific

21/10/2006 20:35:30 Place Specific

12/12/2006 08:29:22 Place Specific

12/12/2006 09:13:30 Place Specific

17/01/2008 13:28:18 Place Specific

28/05/2006 15:45:17 Event Specific

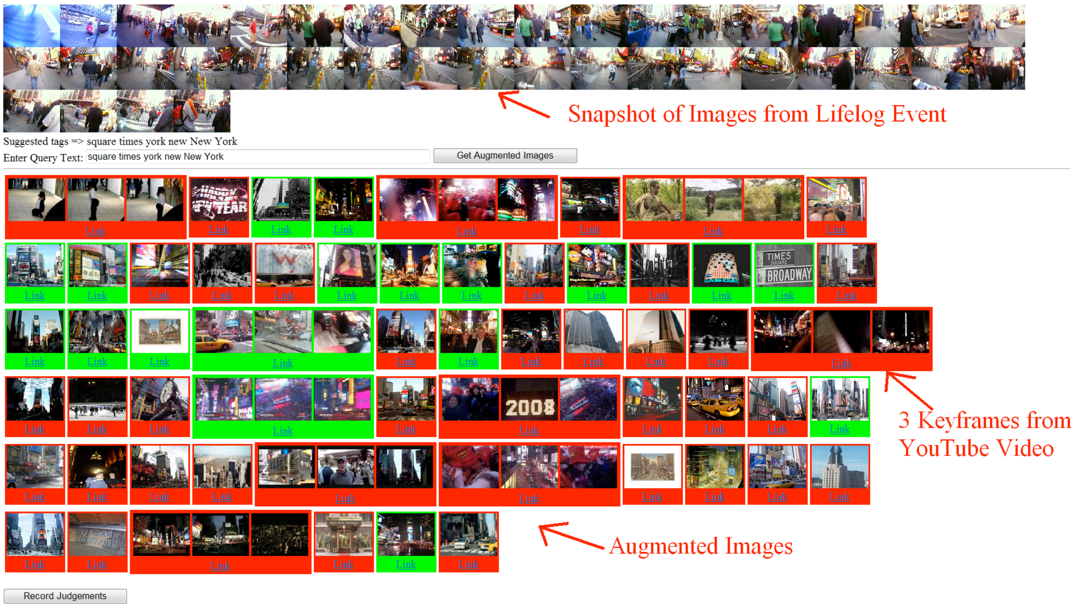
26/10/2006 12:12:55 Event Specific

EVENT DETAIL
(12:10 - 12:18; duration: 7min; 28 images)

Feel free to enter an event description here:

○ Event Specific ● Place Specific
Latitude: 34.4084
Longitude: -119.685
Tag Event

Figure 6.6: A screenshot of the browser used to select events for annotation.



Snapshot of Images from Lifelog Event

Suggested tags => square times york new New York
Enter Query Text: square times york new New York
Get Augmented Images

Figure 6.7: A screenshot of the application used to determine event augmentation relevance judgements.

6.4 Results

This section now discusses the evaluation we carried out on augmenting lifelog events with images/videos from other sources of information. We extensively evaluate the results of each of the three phases of our event augmentation system as illustrated back in Figure 6.2, namely:

- Which source(s) of geo-tagged photos should we retrieve from which to extract user tags?
- How effective is the automatically constructed text query from those tags?
- Do users prefer a fully automated augmentation process, or one where they are suggested a textual query, but given the option to amend that query?

We will now explore our findings in detail.

6.4.1 Which Photos to Retrieve from the Same Location ?

Given that the user will have the time and GPS information of their lifelog event automatically recorded, we retrieve relevant geotagged images from Flickr based on a location search. In describing our approach in a prior section, we posed the question as to whether it is better to retrieve the tags from the top 100 (*T100*) or the top 25 (*T25*) most relevant (in terms of location) geotagged images, or the top 25 geotagged images that are visually similar to the keyframe image from the lifelog event (*M25*).

In practice we found that none of these approaches performed consistently better than any other, although selecting the top 100 geotagged images provided marginally superior results (overall precision of 0.308 vs. 0.190 for top 25 geotagged vs. 0.276 for top 25 visually similar). 34 of all the tagged events had their results generated automatically, and here it was possible to compare the performance of these 3 approaches as displayed in Figure 6.8. Again using the tags from the top 100 geotagged images performs best (0.260

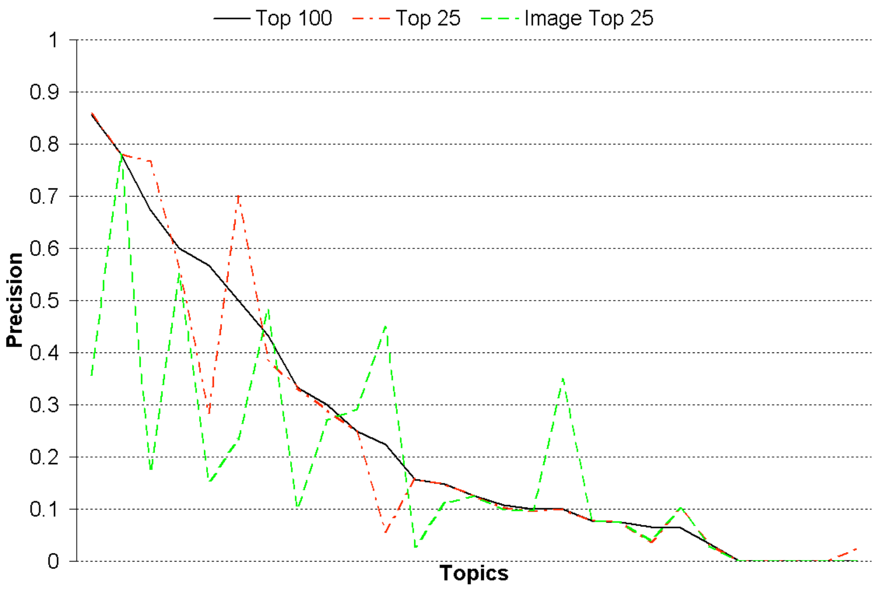


Figure 6.8: An illustration of the event augmentation performance effects in changing the initial seed geotagged photos from which tags can be extracted.

precision)[10]. Selecting the top 25 most visually similar performs worst (0.217), and to compound matters there is a large processing overload associated with this approach by having to extract the MPEG-7 descriptors of the top 100 images returned by Flickr and Panoramio for each event (in practice we have found it takes 10 × the processing time).

We investigated whether the number of seed geotagged images initially retrieved had any impact on the final system performance and in practice we found no such relationship/correlation existed (correlation of -0.09 between number of seed geotagged images and precision of retrieved results).

We discussed earlier that lifelog events, which we would like to augment, can be placed into two broad categories, i.e. whether the lifelog event in question is of an actual place ("place-specific"), or an event or happening such as a big sporting event or rock concert ("event-specific"). So the "place-specific" category can include images taken at any time, whereas the "event-specific" category can only include images taken at a certain time pe-

[10] As no suggested text was presented to the user in automatic runs, users were presented with results from the *T100, T25, & M25* approaches, thus allowing us to compare the retrieval results of these approaches on a like-for-like basis.

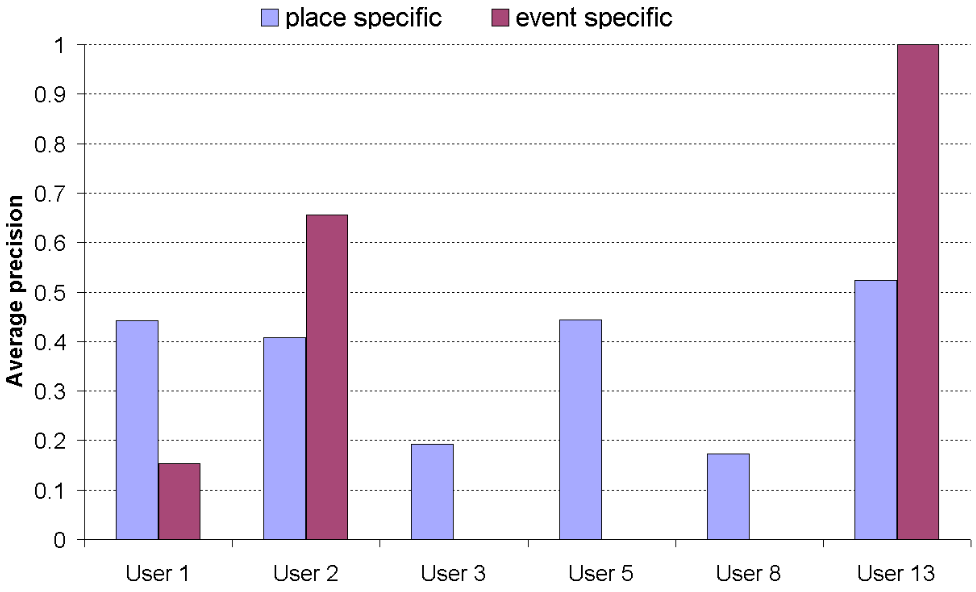


Figure 6.9: The event augmentation differences in the “Place Specific” vs. “Event Specific” results by user.

riod. Overall the “place-specific” event augmentation performed better than the “event-specific”, with an average precision score of 0.305 vs. 0.186. Also out of the 6 users who had both place- and event-specific events, the place specific results were better for 4 of them (see Figure 6.9). The “place-specific” also performs better on both the automatic (0.249 vs. 0.123) and semi-automatic systems (0.448 vs. 0.378). Comments made on this aspect of the augmentation process in our post evaluation questionnaire were concentrated on criticising the “event-specific” results, with one user (User 5 in Figure 6.9) commenting that *“. . . for location-specific events, most of them were good. For event-specific events it was all incorrect . . . ”*. No comments were made when the event-specific system worked better, perhaps indicating an expectation that it should just work anyway.

Given that it is possible to automatically search for geotagged images, we posed the question as to how many images should be retrieved, so as to construct a new text query afterwards. We have found that taking the 100 most spatially relevant images performs best.

6.4.2 Benefits Offered in Using Relevance Feedback to Construct Text Query

After automatically retrieving relevant geotagged images, we proposed earlier that intelligently constructing a new textual query from the tags associated with those retrieved images would be beneficial, and provide the user with many more relevant augmented images/videos. This is due to the fact that while there are now tens of millions of geotagged images available, it is still a small percentage of the *total* amount of potentially relevant images/video available. Thus by being able to construct a textual query, it is then possible to access those other relevant images/videos.

It is indeed the case that constructing a new textual query is beneficial, as 55.15% of the total number of relevant images/videos across all the users on the automatic runs came from the sources of information that depend on the automatically constructed query. This immediately offered users 123% more relevant augmented images.

Considering that users were given the ability to amend the textual queries on the semi-automatic system runs we can use another method to evaluate how closely the suggested text matches the user query. Heavy editing would negate the advantages of automatically suggesting any keywords. To do this we can calculate the number of overlapping words between the suggested text and the actual query input by the user which was taken as the gold standard. This enables us to calculate precision (the total number of system suggested words divided by the number of overlapping words), recall (the total number of actual words input by the user divided by the number of overlapping words), and the F1-Measure [11].

Again using the top 100 geotagged images as a source of tags works best with a precision of 0.489, a recall of 0.582, and an F1-Measure of 0.532. This would indicate that the suggested tags were helpful to the users. However, given that each query is different, and some are much more difficult than others, there is naturally a large variation of scores between users on how accurate the suggested text was (ranging from an F1-Measure of 0.76 to 0.31) as illustrated in Figure 6.10. It is also interesting to notice the trend between the F1-Measure (thick continuous line) and the normalised user Likert rating for the tag sug-

[11] F1-Measure = $\frac{2*precision*recall}{precision+recall}$

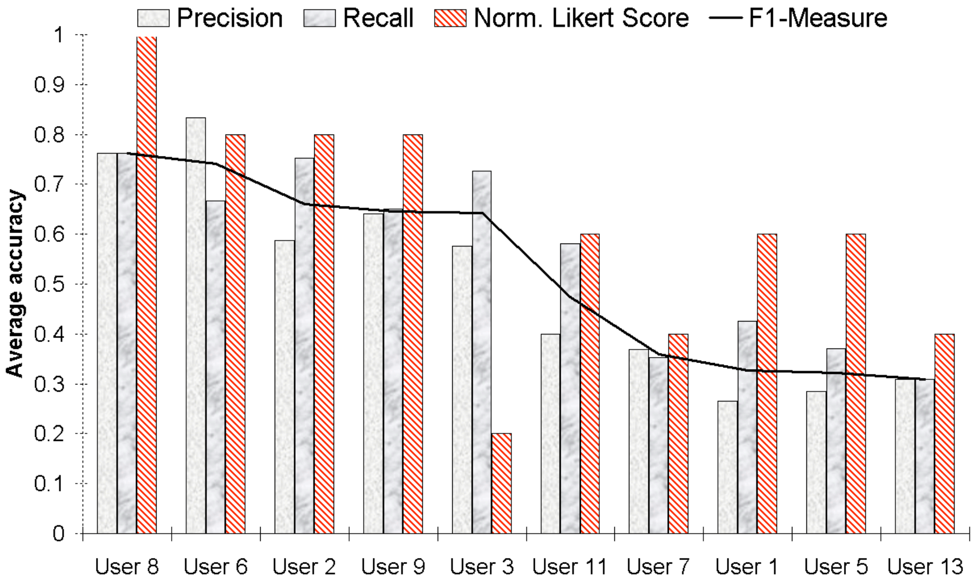


Figure 6.10: An insight into the effectiveness of the tag suggestion for each user on the completely automated runs.

gestion usefulness in the post evaluation questionnaire (diagonally filled column). Broadly speaking these two data series mirror each other (with a correlation of 0.55), thus indicating that users naturally are happier when more accurate query text is suggested. User number 3 was the only person to buck this general trend, feeling that the suggested text was not sufficiently specific, *"...You really need a local knowledge of the area to form the query correctly ... "*.

Indeed, in a post-evaluation questionnaire there were some negative comments in which the most common complaint was that the textual suggestions were not "specific" enough, e.g. provides the city/street name, but not the name of the particular building that the user was interested in. However, in total 7 out of the 10 users have found the tag suggestions helpful in many instances[12]. The median Likert score of all 10 users was 3/5.

[12] User number 4 from Table 6.3 only had 1 query, which was determined to be a completely automated run by our Latin Squares setup. Therefore only 10 users instead of all 11 could be evaluated in this section

6.4.3 Augmented Image/Video Search Results

In this chapter we have introduced the concept of augmenting lifelog events with publicly available multimedia content that has been generated by millions of other users. Such content is available from a number of different data sources, and we now comment on those sources that consistently provide a greater concentration of relevant material to users.

As has been previously mentioned, we feel that due to the relatively small percentage of publicly available geotagged content, it will be useful to construct a new text query based on tags associated with retrieved images, so as to search for *many* more potentially relevant images/videos to a user's lifelog events. Having automatically constructed a new text query on which to search for other sources of information (e.g. Yahoo!, YouTube, MSN, etc.), we proposed two systems to the user: 1) All the images are automatically returned to the users; and 2) The suggested text is shown to the user (who can edit it) before retrieving potentially relevant results.

The semi-automatic system's retrieval results were better for 9 out of the 10 people who used both systems as illustrated in Figure 6.11 and Figure 6.12. In fact the overall precision of the semi-automatic system was almost twice that of the automatic system (0.441 vs. 0.235), thus indicating that user feedback is vital in locating relevant images. In addition to the superior retrieval performance, 9 of those 10 users also simply preferred using the semi-automatic system as it provided them with more security, with one user eloquently summing up the feelings of others by stating that *"...I missed the textbox when it was removed ... "*.

Considering that the top 10 results are returned to the users from 6 sources of information[13], we now report those sources that provided the highest concentration of relevant results. Investigating all judgements made, searching Flickr by text had the highest average precision score (0.370) followed by MSN (0.329), Yahoo! (0.290), Panoramio geotagged images (0.280), Flickr geotagged images (0.242), and finally YouTube (0.205) as illustrated

[13] For geotagged images: Flickr and Panoramio; For text search images/video: Flickr, Yahoo!, MSN, YouTube

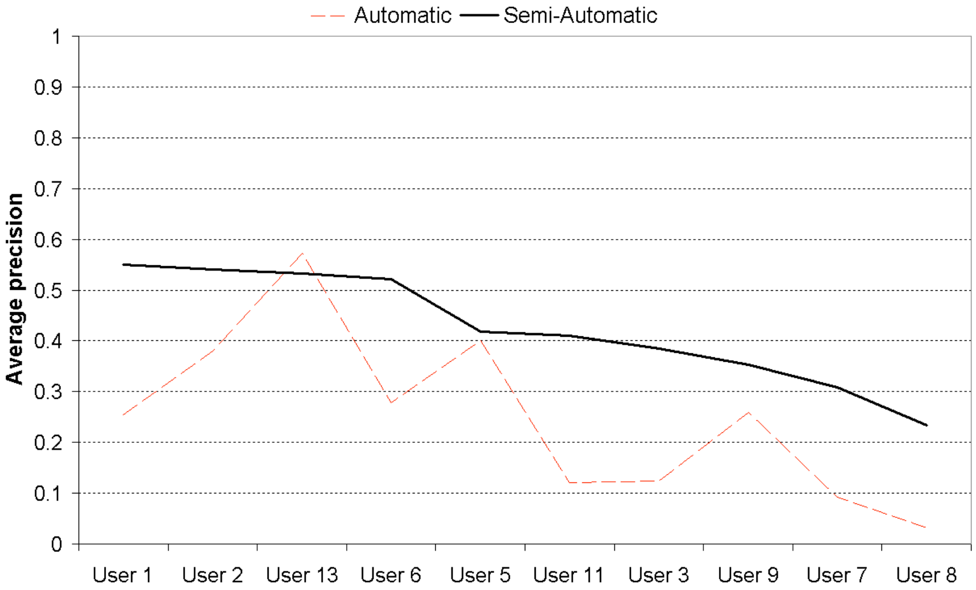


Figure 6.11: The effectiveness of the (semi-) automated runs on each user for event augmentation.

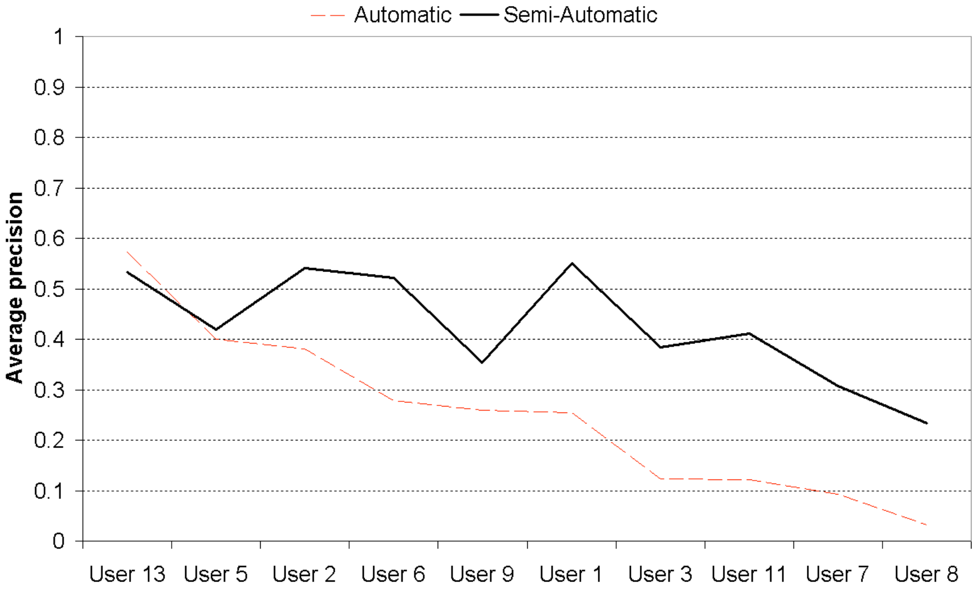


Figure 6.12: The effectiveness of the automated runs on each user for event augmentation.

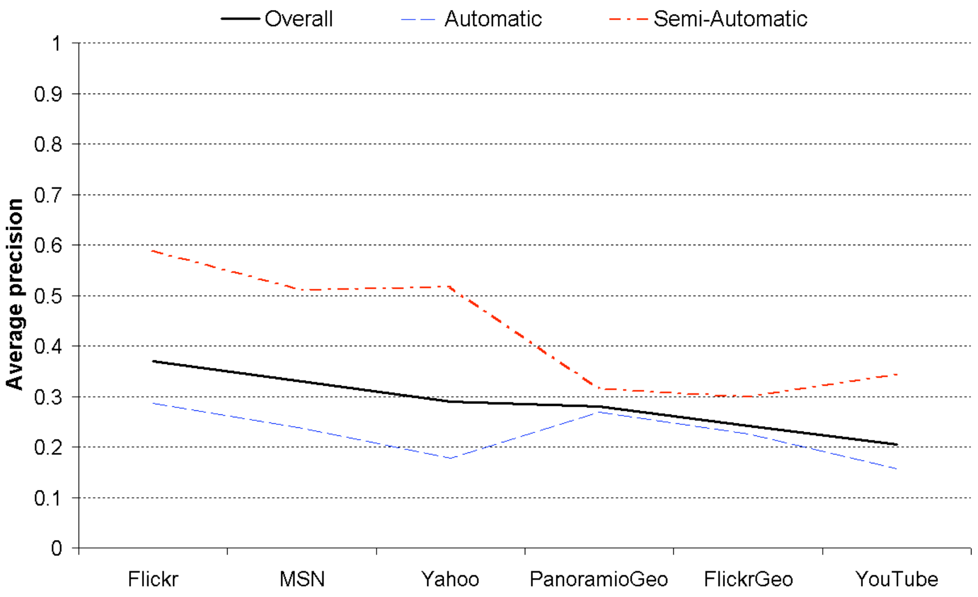


Figure 6.13: An insight into the final number of relevant results returned by each external source of information.

in Figure 6.13. Given that users much prefer the semi-automatic system, it is interesting to see which sources of information provided the highest number of relevant results on semi-automatic runs (the medium thickness dot dashed line in Figure 6.13). Flickr text search again works best but this time with a much higher precision score of 0.590; Yahoo! is next best (0.516), then MSN (0.510), then YouTube (0.343); and finally the two geotagged sources of Panoramio (0.315) and Flickr (0.299). So what is really interesting to note here is that on semi-automatic augmentation runs, which users prefer anyways, the "text only query" sources clearly perform better.

In this section various facets of our approach have been evaluated in augmenting lifelog events with images/videos from other sources of information. Now we investigate how effective for users the optimal combination of these approaches is. Based on these results, the best overall system would use the top 100 retrieved geotagged images as a source from which to construct a text query. The user would then be given the opportunity to amend the generated text query, and the Flickr (text), MSN, and Yahoo! sources would be used

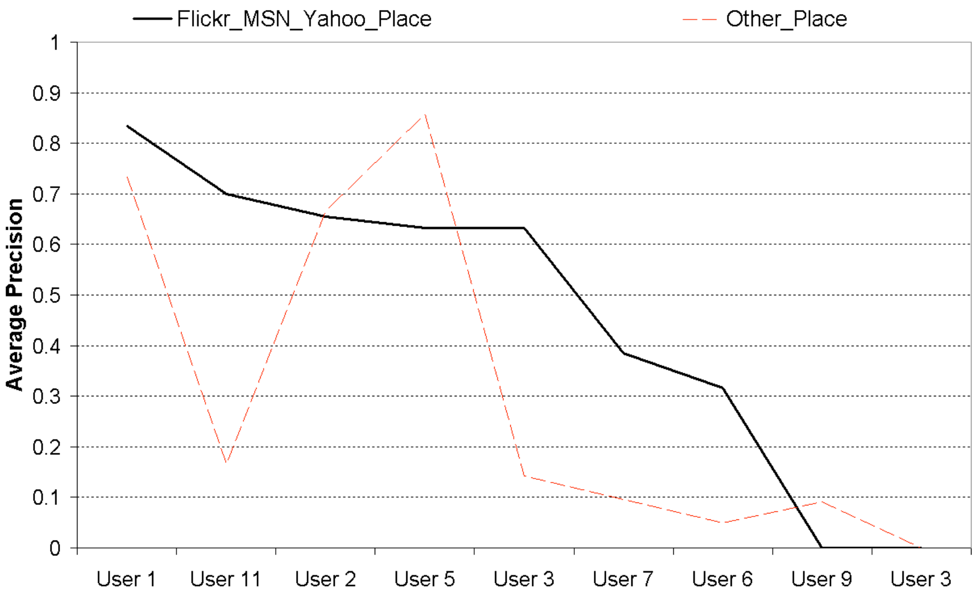


Figure 6.14: An illustration of the performance of the final recommended system on “place-specific” events. The top 100 most spatially relevant images are firstly retrieved by Flickr, our technique then automatically selects 5 tags, which the user can amend (semi-automatic), before retrieving images/videos based on text-only search.

to present augmented image candidates to the user. Of the 9 “place-specific” runs that had these parameter settings in our experiments, a median precision score of 0.633 was recorded, meaning that users found two-thirds of the presented images useful for augmenting their lifelog events, as illustrated by the thick continuous line in Figure 6.14. It is also possible to include the YouTube, Panoramio and Flickr (geotagged) results, but there will be many noisy/irrelevant results included as these 3 sources had an average precision score of just 0.350 on the 9 aforementioned runs (shown on the dashed line in Figure 6.14).

There were only 2 “event-specific” runs (with the optimal combination of facets) that could be evaluated in our experimental dataset, therefore it is difficult to draw any meaningful conclusions. However, on these two runs, only one returned any results at all (0.4). This indicates that in future there exists a significant challenge in terms of augmenting “event-specific” lifelog events with relevant content. On inspection of Table 6.2 from quite a few pages back in section 6.2.3 it can be seen that “places of interest” have many more relevant images (e.g. Sugar Loaf, Big Ben, etc.) than “events of interest” (e.g. soccer world cup final, AFL grand final, etc.). This means that there is a smaller number of potentially relevant results available which makes the challenge more difficult than retrieving relevant "place-specific" queries.

6.5 Summary

In trying to provide as many memory retrieval cues as possible, one of the aims in our hypothesis is to *"... augment images from the low-resolution wearable device with higher quality images from external data sources ... "* The aim of the work reported in this chapter was to thoroughly investigate the idea of augmenting passively captured lifelog events with relevant image/videos captured by a large number of individuals, each of whom only had to make a small contribution. In this chapter we have also noted the phenomenal expansion of easily-accessible multimedia content due to "Web 2.0", whereby millions of users are uploading billions of images and videos. We are now entering an era whereby individuals expend little effort in capturing images and making them available, whether it is passively captured lifelog images or a small number of manually captured photos which are uploaded to a media sharing website. However, in aggregating these small contributions over an enormous scale of users we have automatically enriched the experience of individuals reviewing their trip or event, by providing them with a large number of relevant items of information mined from millions of other individuals.

We have proposed techniques that make possible this goal of augmenting lifelog events. Given that lifelogging devices can easily capture time and location information, it is possible to automatically search for "geotagged" images on the Internet. However, as we have discovered in this chapter, very little of the total number of potentially relevant images/videos on the Internet are actually geotagged. However, all of this content can be retrieved via textual queries. Therefore in this chapter we introduce a technique to initially (and automatically) retrieve relevant geotagged images, and then to construct a textual query by examining the tags associated with those retrieved geotagged images i.e. this is a form of pseudo relevance feedback. Using that textual query it is then possible to retrieve potentially relevant images/videos from *many* sources of information on the Internet.

To investigate the effectiveness of such a system we have organised 11 users to collect SenseCam data over a two year period. 1,937,770 images were captured in total, which were automatically segmented into 22,911 events. The users were then asked to identify a selection of events on which they would like to see augmented material, and in total 67 such events were selected, 56 of interesting places and 11 of interesting events. Afterwards each user was presented with potentially relevant results, as determined by various evaluated approaches, and in total 4,963 judgements were made on augmented images/videos for the 67 selected events.

We evaluated a number of different techniques in the results section and the first finding is that it is optimal to initially retrieve the 100 most spatially relevant geotagged images to the lifelog event in question. Not only is it quite accurate to select the most spatially relevant images, but it is also approximately 10 times quicker, in terms of processing time, than selecting by visual similarity.

It was found that constructing a new textual query is very beneficial as 55.15% of the total number of relevant images/videos across all the users on the automatic runs came from the sources of information that depended on the automatically constructed query. This immediately offers uses approximately 120% extra relevant images/video to augment their experience of reviewing a passively captured lifelog event. In a post-evaluation questionnaire we discovered that 7 out of 10 users found the suggested text quite helpful too.

In the evaluation stage the users were presented with two system variations: 1) Where the results were automatically generated with no user intervention, 2) Where the textual query was automatically generated, but users were given the opportunity to amend it before searching for relevant content (referred to as *semi-automatic* approach). In practice it was found that 9 out of the 10 users who were presented with both systems preferred the semi-automatic system. In fact the semi-automatic system returns twice as much relevant content to the user, (0.441 P@10 vs. 0.235), thus indicating that user feedback is vital in locating more relevant images.

Potentially relevant content from 5 different sources of information was evaluated: Flickr, Panoramio (Google Earth), MSN, Yahoo!, and YouTube. It was found that MSN, Yahoo!, and searching Flickr by text were the richest sources of information in terms of yielding relevant results, especially on the semi-automatic system runs. Overall it was found that running a system with the optimal combination of facets on “place-specific” runs, presents the users with approximately 18 relevant images out of 30 (from MSN, Yahoo!, & Flickr with each information source providing 6/10 relevant results per query). In our experiments only 2 ”event-specific” runs could be evaluated with the optimal combination of facets, and naturally no concrete observations can be made, however it does appear that these present a considerable future challenge.

Even though there are research challenges involved in further improving the quality of the lifelog augmentation process, we have now made real one of the important goals of pervasive user-generated content, as an individual’s passively captured content in their lifelog is augmented with user-generated content collected by millions of other individuals.

6.6 Conclusions and Contributions

- Given that only a small percentage of photos uploaded to “Web 2.0” sites are geotagged, we have illustrated that it is necessary to search for relevant content through text.

- We have introduced an approach to initially select some seed images through an automated spatial search, and thereafter to construct a text query from the tags associated with those seed images.

- In selecting a set of seed images we have discovered that it is not worthwhile to identify tags from those images that are visually similar to the keyframe image of the given SenseCam event.

- A future area of interest to explore will be that of selecting the most representative tags from the seed geotagged images. An adaptation of the TF-IDF algorithm [79] in particular may be worth looking at.

- Another future area of interest may be exploring the usefulness in automatically annotating one's lifelog events through the tags produced in augmenting one's event.

Chapter 7

REVIEWING HYPOTHESIS THROUGH ANECDOTAL EVIDENCE

The previous chapters in this thesis have concentrated on evaluating the accuracy of individual facets of the hypothesis of this dissertation in a very systematic format. Results have focused on maximising the F1-Measure, precision, recall, or mean average precision scores. However, we have still not investigated if facets such as event segmentation are *really* of benefit to users in an end-system. Something as potentially fulfilling as reflecting on one's personal experiences can not only be reduced to precision and recall numbers. Therefore in this chapter we carry out usability tests to explore how fun, insightful, and most importantly how useful each facet of the system is to real-world users. In previous chapters we have alluded to the fact that there exists a body of literature that points towards the fact that cued-based recall is better than free recall, and that images in particular provide strong retrieval cues. In effect we are now testing the hypothesis of this dissertation to investigate its usefulness in providing effective memory retrieval cues. To recap the hypothesis of this dissertation is that:

In order for visual lifelogging technologies to effectively provide autobiographical memory retrieval cues it is necessary to automatically:

- *effectively segment a large collection of personal images into distinct events*
- *accurately determine which of those events are most unusual*
- *usefully provide the ability to link similar events together*
- *augment images from the low-resolution wearable device with higher quality images from external data sources.*

It is beyond the scope of this thesis to carry out large scale user tests, as this would require many users gathering large amounts of data over a long period of time. An ethnographic study of this nature would require a very significant body of work. Instead in this chapter we will investigate, in an anecdotal manner, the performance of each individual aspect e.g. event segmentation, retrieving similar events, keyframe selection, determining event importance, and event augmentation. We present two distinct systems, one an event-based browser that we distributed to international research groups, and the other that incorporates all facets of the visual lifelog processing cycle.

Before we discuss the findings of our discussions with users, we will quickly recap on the various aspects of the system that are being evaluated. Thereafter we will discuss in detail two browsers that we used to conduct usability tests on those various lifelogging facets. Finally we will detail our findings based on feedback offered by a number of users.

7.1 Aspects to Investigate

We will now provide a brief reflection on the various approaches discussed throughout the duration of this thesis which relate to individual facets of the hypothesis in this dissertation. The following approaches (illustrated in Figure 7.1) have already been quantitatively investigated already, but will now be looked at in an anecdotal manner in this chapter:

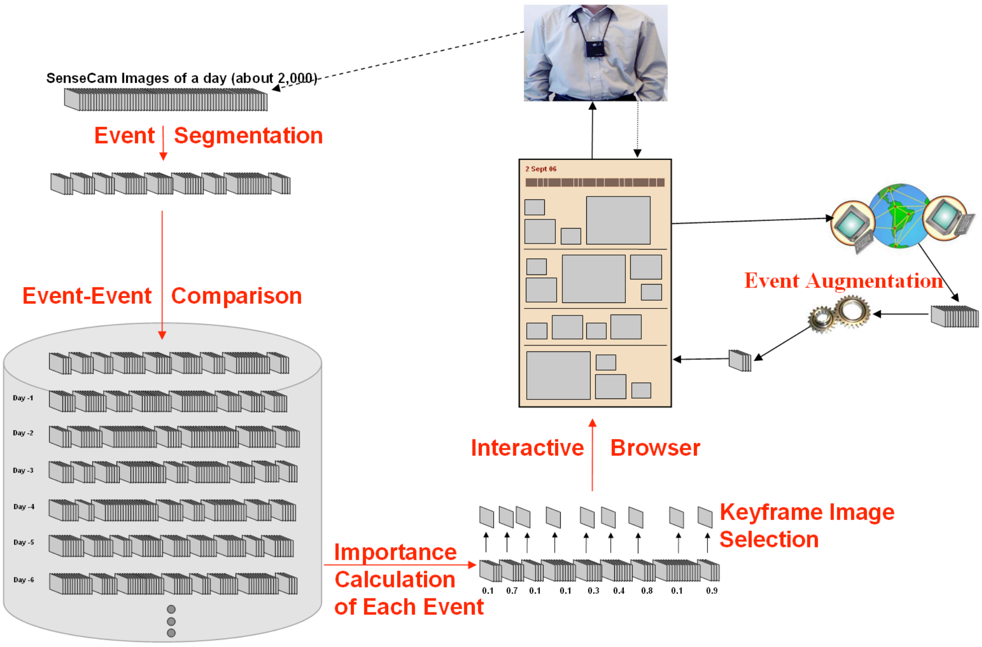


Figure 7.1: An overview of the challenges that this thesis has addressed thus far. Images are segmented into distinct events, which then allow those events to be compared against each other. By comparing events against each other we can then determine an importance score for each and every event. A keyframe must be selected for each event, which can also be augmented with material from external sources of information. In this chapter we will review a system incorporating the outputs of the prior chapters.

- **Event Segmentation:** Chapter 3 details many quantitative experiments on the optimal approaches towards segmenting sequences of SenseCam images into distinct events or activities. Examples of events/activities include: having a meal, being on a bus/train, talking to a friend, going for a walk in the park, etc. In essence event segmentation algorithms attempt to identify the boundary between each event, which occurs at times of high visual or sensory changes. The recommended approach from Chapter 3 is performed on only the sensor sources of information recorded by the SenseCam. This results in very quick segmentation processing.

- **Event Searching:** Chapter 4 provides details on many quantitative experiments on approaches towards the retrieval of similar/relevant SenseCam events to a given lifelog event. This may be very useful when a user is trying to find other times that they went to the park, or other times they had a meal in a restaurant, or other times that they were driving, etc. Event retrieval algorithms must first correctly represent each event (as events consist of almost 100 images on average), and then identify other events that are close both visually and in terms of sensor readings.

- **Keyframe Image:** Chapter 5 provides details on experiments towards selecting an appropriate keyframe image from an event. As lifelog events consist of almost 100 images on average, a significant challenge exists in selecting an image that represents the semantic meaning of an event. This image can then be used in an end system browser to allow users to recognise events based on just 1 image rather than 100 images! In Chapter 5 we recommend selecting the image within the event that has the highest "quality" score.

- **Event Importance:** Chapter 5 also details extensive experiments towards calculating an importance value for every event in a lifelog collection. On average users will have over 20 events per day, thus we feel it will be useful for any real-world application to place more emphasis on those events that are more interesting/important. We determine importance based on 2 aspects: 1) the concentration of detected faces in the event (thus indicating social interaction) and 2) the uniqueness of the event with respect to all other events in the past fortnight.

- **Event Augmentation:** Chapter 6 provides details on augmenting SenseCam events with images from other sources of information. Given the prevalence of "Web 2.0" sites on the Internet nowadays, and the massive number of publicly available geo-tagged images, it is now possible to automatically query such sites for potentially relevant content to augment a user's experience of reviewing a particularly event, e.g. a trip to the Eiffel tower in Paris, the Colosseum in Rome, or to a big rock concert. In this chapter we will investigate if this aspect is of real benefit and enjoyment to users in an actual system.

7.2 Experimental setup

To evaluate the hypothesis of this dissertation we presented 9 users with two different systems to review their own data. One of these systems is an event based browser which was distributed to four collaborating universities: the University of Tampere, Finland; Duke University, USA; University of Toronto, Canada; and Utrecht University, in the Netherlands. The other system is one used in-house in the Centre for Digital Video Processing in Dublin City University, and this system incorporates all the features listed in the previous subsection: event segmentation, event searching, keyframe image selection, determining event importance, and finally augmenting events with images from external sources of information. We will now detail both of these systems and the users they served.

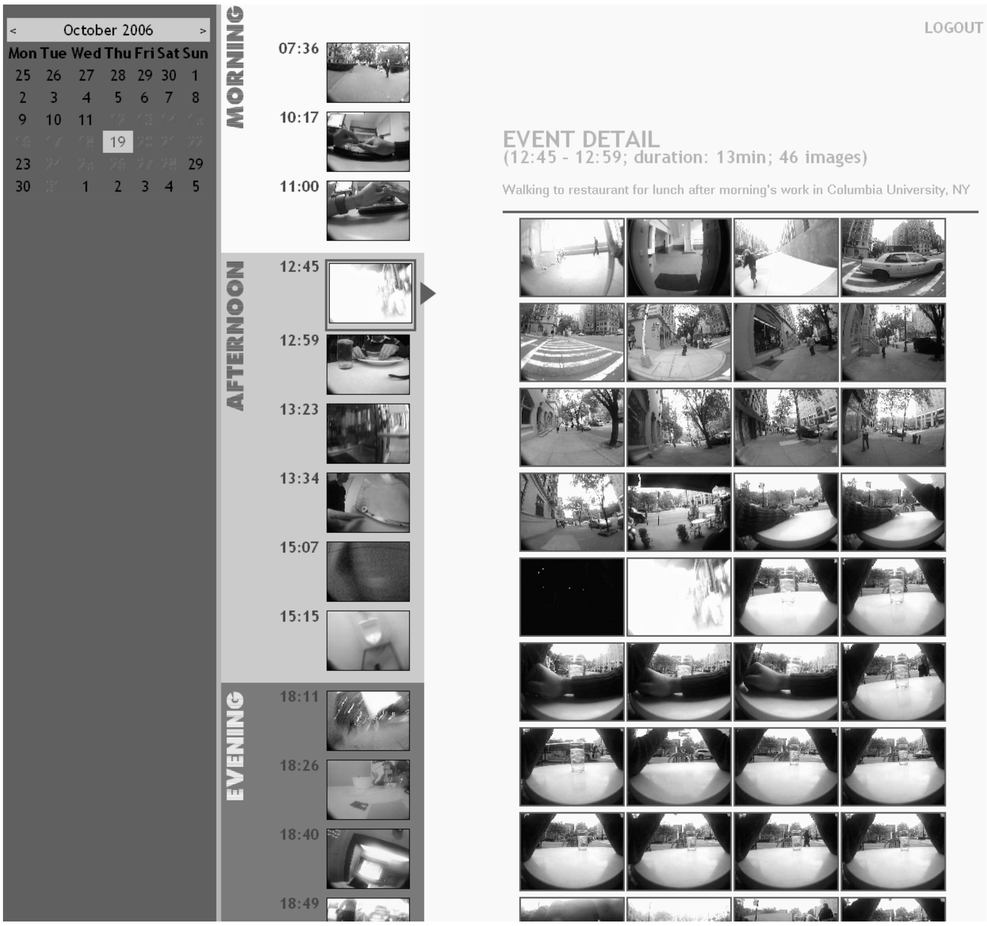
7.2.1 Event-Based Browser

We have worked with four universities (Tampere (Finland), Duke (USA), Toronto (Canada), and Utrecht in the Netherlands) and provided them with SenseCams and also an event-based browser. We have plans to work with other research institutions who have expressed an interest in using this browser in the future too. This browser first of all automatically segments sequences of collected images into distinct events or activities, based on the sensor values alone. These events are then visualised as displayed in Figure 7.2. The top left corner of the screen is a daily calendar which allows a user to browse to a desired day, exploiting the fact that humans can remember the approximate time that events take place, as noted in Chapter 2. The currently selected day in the calendar is highlighted by a different background colour (the 19^{th} of October 2006 in Figure 7.2). The middle column of the browser displays a single keyframe image of every event present in the selected day. The middle image in each event is selected as the keyframe image, meaning processing is very quick. The events for the morning, afternoon, and evening have different background colours, so as to allow the user navigate to the section of the day in question that they are most interested in reviewing. Events that are selected to be browsed in detail are highlighted with a red border and arrow as illustrated with the first "Afternoon" event in Figure 7.2. The right hand side of the browser then displays all the individual images present in a selected event. Above the event's images, other details are given about the event, e.g. the start/end times, the duration of the event, and also any associated comment. Users can add any comments to an event. Finally if the user is interested in any particular image within the event they can click on it to display it in full size as illustrated in Figure 7.3.

The event-based browser was used by one user in Utrecht to review his own personal lifelog images. This user collected 102,113 images over a period of 55 days which were segmented into 1,290 events. The evaluation section will focus on how effectively the user's images were segmented into distinct events/activities.

In Tampere, the event-based browser was used for a different purpose, whereby one user was reviewing images received from 6 individuals who wore the SenseCam over a period of 10 days each. The purpose here was to use the SenseCam as a data collection tool in an empirical study on information access in molecular medicine [29]. In total the 6 individuals collected approximately 72,000 images over a period of 52 days[1] which were segmented into approximately 800 events. The evaluation section though will focus on

[1] One of the 6 individuals only captured 2 days of data



< October 2006 >

Mon	Tue	Wed	Thu	Fri	Sat	Sun
25	26	27	28	29	30	1
2	3	4	5	6	7	8
9	10	11	12	13	14	15
16	17	18	19	20	21	22
23	24	25	26	27	28	29
30	31	1	2	3	4	5

EVENT DETAIL
(12:45 - 12:59; duration: 13min; 46 images)

Walking to restaurant for lunch after morning's work in Columbia University, NY

Figure 7.2: A screenshot of the "event-based" browser.

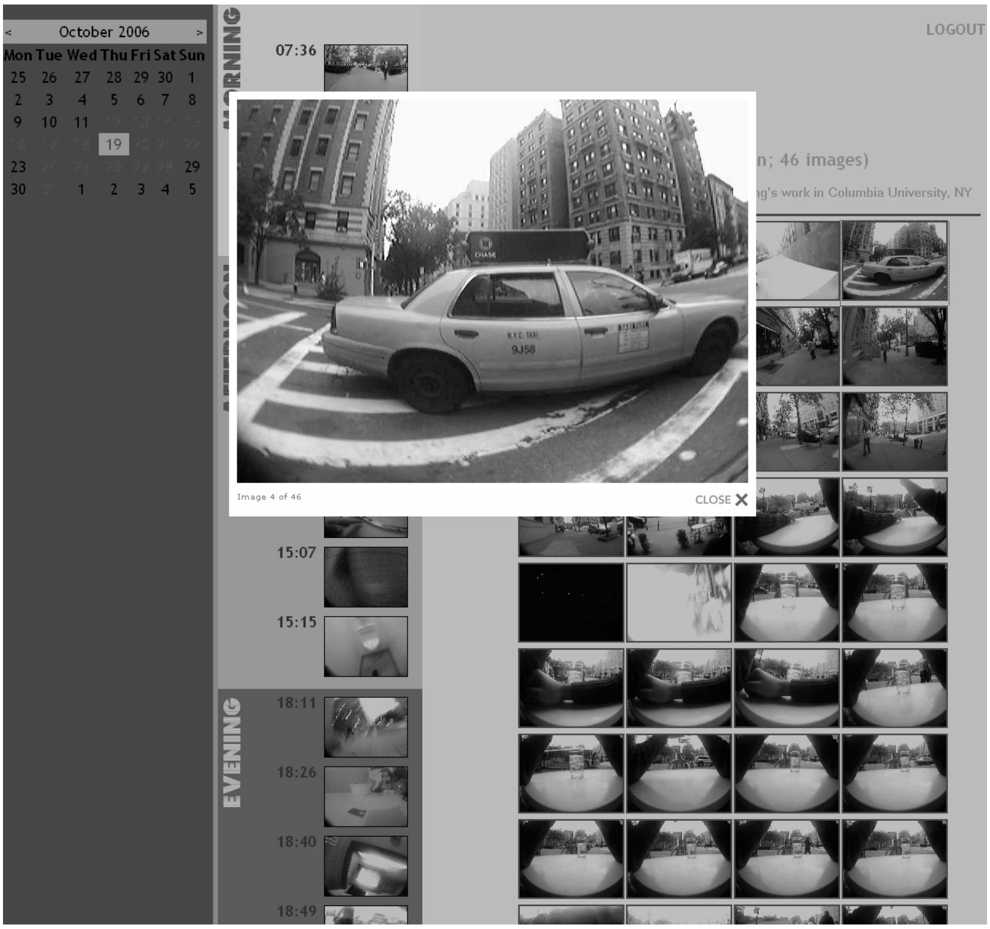


Figure 7.3: Displaying a single image on the event-based browser.

feedback from just one person (the information researcher), and how useful she found the event-based browser to determine the information needs of the molecular medical workers, with a particular focus on how effective the event segmentation is.

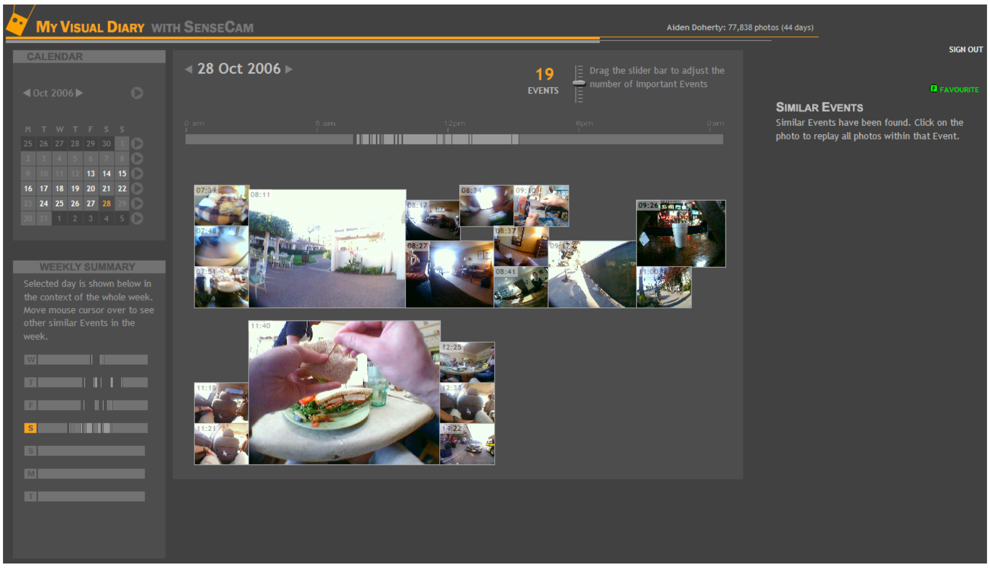
In Toronto, the event-based browser is used as part of a project to enhance the quality of life of people with Alzheimer's through using SenseCam images. 2 people have used this browser to investigate its suitability towards providing memory retrieval cues. One of the users collected 5,524 SenseCam images over a period of 5 days which were segmented into approximately 67 events, with the other user collecting a much larger dataset. The evaluation section will focus on how effectively their images were segmented into events.

Meanwhile in Duke University, the event-based browser is used as part of a project to investigate why and how SenseCam images provide such powerful memory cues. 1 person has used this browser to review images captured by subjects, so as to step through their days in a collaborative fashion. Our evaluation section will focus on how effective this user feels that our event-based browser is.

7.2.2 Advanced Browser

We now present a SenseCam browser that incorporates all of the facets detailed in our hypothesis: event segmentation, event searching, keyframe image selection, determining event importance, and finally augmenting events with images from external sources of information. We will detail how these aspects are visualised on our browser.

Figure 7.4 provides an overview of our browser. The top left region of the browser provides the user with calendar functionality where the user can select any desired day to show. Again, as with the event-based browser this exploits the fact that people remember the approximate time that events take place. Days available for selection are highlighted in bright white text, with the current day on view highlighted via yellow font. The predominant focus of the browser is on the middle of the screen where numerous images of the day are displayed. These images were segmented into distinct events/activities, and on the browser each image is an event keyframe image, i.e. the image with the highest quality score.



M	T	W	T	F	S	S
25	26	27	28	29	30	1
2	3	4	5	6	7	8
9	10	11	12	13	14	15
16	17	18	19	20	21	22
23	24	25	26	27	28	29
30	31	1	2	3	4	5

Drag the slider bar to adjust the number of Important Events

Selected day is shown below in the context of the whole week. Move mouse cursor over to see other similar Events in the week.

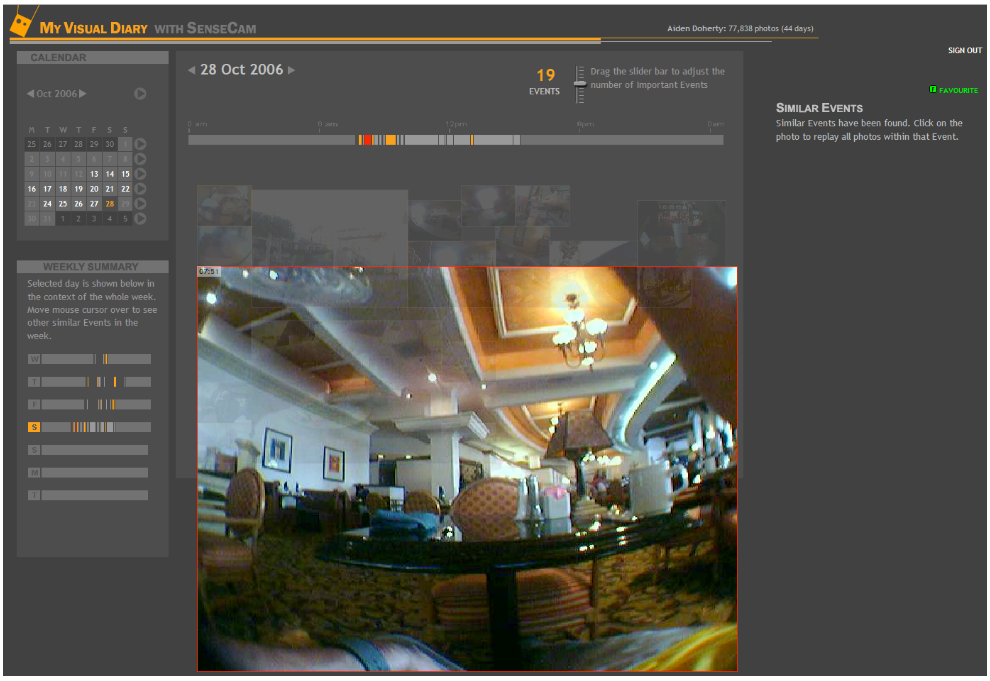
Similar Events

Similar Events have been found. Click on the photo to replay all photos within that Event.

Figure 7.4: A screenshot of the "advanced" browser.

As can be seen, some images are larger than others, whereby those images that are larger correspond to events with a high importance score, suggesting more interesting events to review which exploits the human memory characteristic whereby distinct events are more strongly encoded. On the top left corner of each image, its starting time is displayed. The information at the top centre of the browser details the day being displayed and also the number of events being displayed. The slider bar can adjust the number of events being displayed, and in the case of Figure 7.4 the 19 most important events, as determined by the system, are displayed to the user. Below the slider bar there is a timeline visualising when events started, and also their duration. The panel to the left of the browser titled "Weekly Summary", just below the calendar section, provides a timeline of other events in the days surrounding the day being displayed. Only those events that are similar to any of the events in the displayed day, are visualised in the weekly timeline, taking advantage of the fact that the human mind stores items associatively, and thus is curious to see other related events.

Figure 7.5 illustrates what happens when the user puts the mouse cursor over any event in the centre of the screen, whereby the images are displayed at full size and played back at a rate of 10 frames per second. The playback is stopped when the user exits the mouse



M	T	W	T	F	S	S
25	26	27	28	29	30	1
2	3	4	5	6	7	8
9	10	11	12	13	14	15
16	17	18	19	20	21	22
23	24	25	26	27	28	29
30	31	1	2	3	4	5

Drag the slider bar to adjust the number of Important Events

Similar Events

Similar Events have been found. Click on the photo to replay all photos within that Event.

Selected day is shown below in the context of the whole week. Move mouse cursor over to see other similar Events in the week.

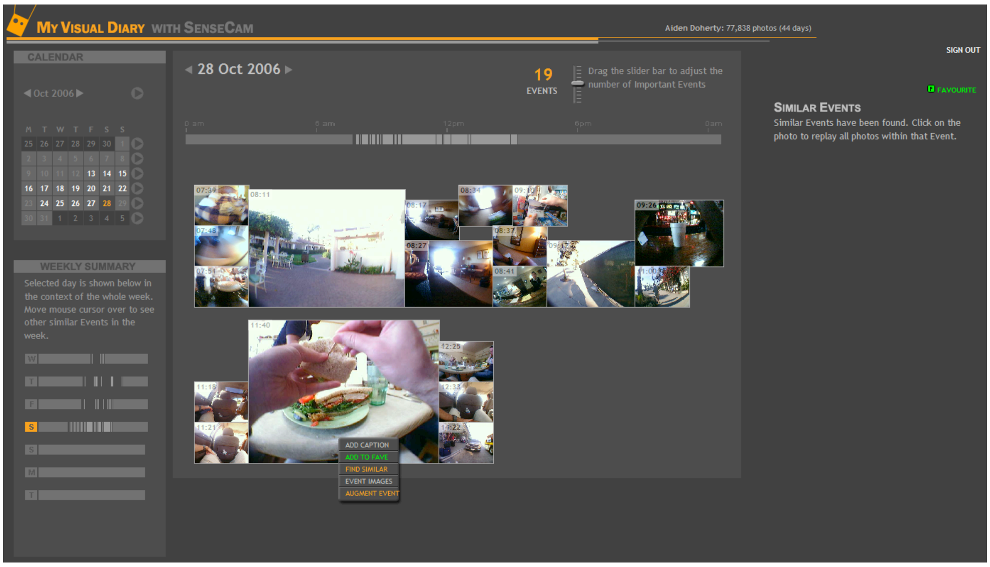
Figure 7.5: The event playback feature on the advanced browser - notice event images are enlarged and played at 10 frames/second.

cursor from the event. The other events are still visible and transparent in the background. Also the event in question is highlighted in red in the daily timeline above, and the weekly timeline to the left too. Any events that are similar to the event in question are highlighted in amber on the daily and weekly timelines.

Thus far we have displayed aspects of the application that highlight the usefulness of: event segmentation in terms of detecting the different events; automated event importance calculations to emphasise more interesting events; and of selecting a good keyframe for visual display. However, there are other capabilities that are useful to provide to the user. As such, we provide the user with a menu of options anytime that they click on any event being displayed, as illustrated in Figure 7.6. This then empowers the user with much extra functionality.

The user can add/amend a textual comment to any event via selecting the "Add Caption" option on the menu, as illustrated on the bottom centre of the browser in Figure 7.7.

The user can also add an event to their list of favourite events via selecting the "Add

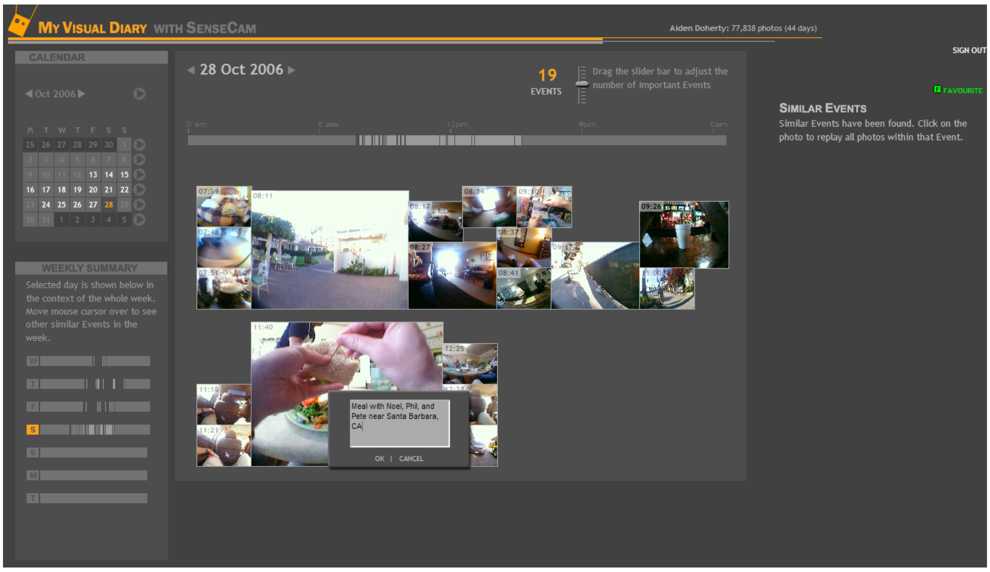


Drag the slider bar to adjust the number of Important Events

Similar Events have been found. Click on the photo to replay all photos within that Event.

Selected day is shown below in the context of the whole week. Move mouse cursor over to see other similar Events in the week.

Figure 7.6: Event menu options on advanced browser - note the floating menu over event.



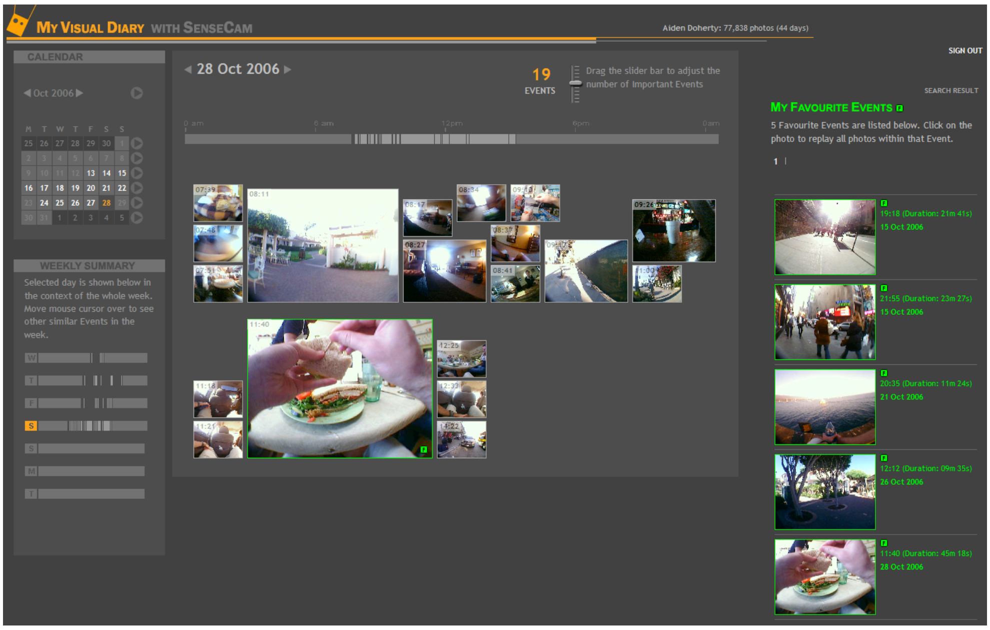
Drag the slider bar to adjust the number of Important Events

Similar Events have been found. Click on the photo to replay all photos within that Event.

Selected day is shown below in the context of the whole week. Move mouse cursor over to see other similar Events in the week.

Meal with Noel, Phil, and Pete near Santa Barbara, CA|

Figure 7.7: Event commenting tool on advanced browser - note the user can enter text in floating text box to comment on event.



Drag the slider bar to adjust the number of Important Events

Selected day is shown below in the context of the whole week. Move mouse cursor over to see other similar Events in the week.

5 Favourite Events are listed below. Click on the photo to replay all photos within that Event.

19:18 (Duration: 21m 41s)
15 Oct 2006

21:55 (Duration: 23m 27s)
15 Oct 2006

20:35 (Duration: 11m 24s)
21 Oct 2006

12:12 (Duration: 09m 35s)
26 Oct 2006

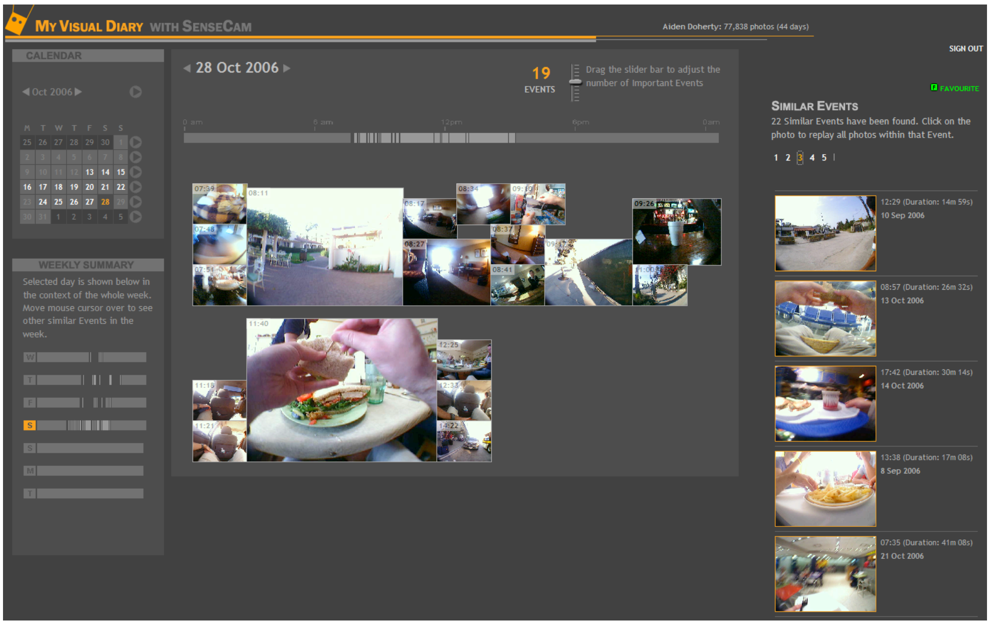
11:40 (Duration: 45m 18s)
28 Oct 2006

Figure 7.8: Note the list of "favourite" events on the right hand panel.

To Fave" option on the menu, as illustrated in Figure 7.8 where the most important event (i.e. the largest one displayed on the browser) has been added to the "My Favourite Events" panel on the right hand side, and also a small "F" caption has been added to the bottom right of the keyframe image of that most important event. Events can also be removed as favourite events too.

Given that the human memory stores items associatively, the user is also provided with the important capability of searching for other potentially relevant events in their lifelog to a given event, by selecting the "Find Similar" option on the menu. A list of other potentially relevant events to the event in question are displayed on the "Similar Events" panel to the right hand side of the browser as illustrated in Figure 7.9, where keyframe images from the list of potentially relevant events to a given event are displayed. Other information regarding the date, time, and duration of the potentially relevant events is displayed too. The top of the panel also gives the user feedback on how many potentially relevant results were returned (22 in the case of Figure 7.9).

As the event playback rate is at 10 frames per second, it is excellent at giving the user



Selected day is shown below in the context of the whole week. Move mouse cursor over to see other similar Events in the week.

Drag the slider bar to adjust the number of Important Events

22 Similar Events have been found. Click on the photo to replay all photos within that Event.

12:29 (Duration: 14m 59s) 10 Sep 2006

08:57 (Duration: 26m 32s) 13 Oct 2006

17:42 (Duration: 30m 14s) 14 Oct 2006

13:38 (Duration: 17m 08s) 8 Sep 2006

07:35 (Duration: 41m 08s) 21 Oct 2006

Figure 7.9: Searching for relevant events on advanced browser - note the list of potentially relevant events on the right hand panel.

the opportunity to "gist" as to what the event is all about. However, if the user is interested in viewing a certain image in any given event, we provide them with the opportunity to view such images in a new tab of the browser as illustrated in Figure 7.10. This new tab is opened when the user selects the "Event Images" option from the floating event menu bar. Once the tab is open, the user is presented with thumbnails of all the images in a given event, and they can then select any image to view it in full size.

Finally given that a user may have attended a certain number of interesting locations, we present them with the opportunity to augment events with images/videos from other sources of information on the Internet to provide them with additional memory retrieval cues e.g. Flickr, YouTube, MSN, Yahoo, or Panoramio (Google Earth). These events must have latitude and longitude information recorded along with them. An event can then be augmented with other images by selecting the "Augment Event" option from the menu, and then the augmented images will be displayed in a new tab on the browser as illustrated in Figure 7.11. Below each result image/video is a "Link" that they can click on to go to the web page where the result originated from, so that they can view the image/video of interest

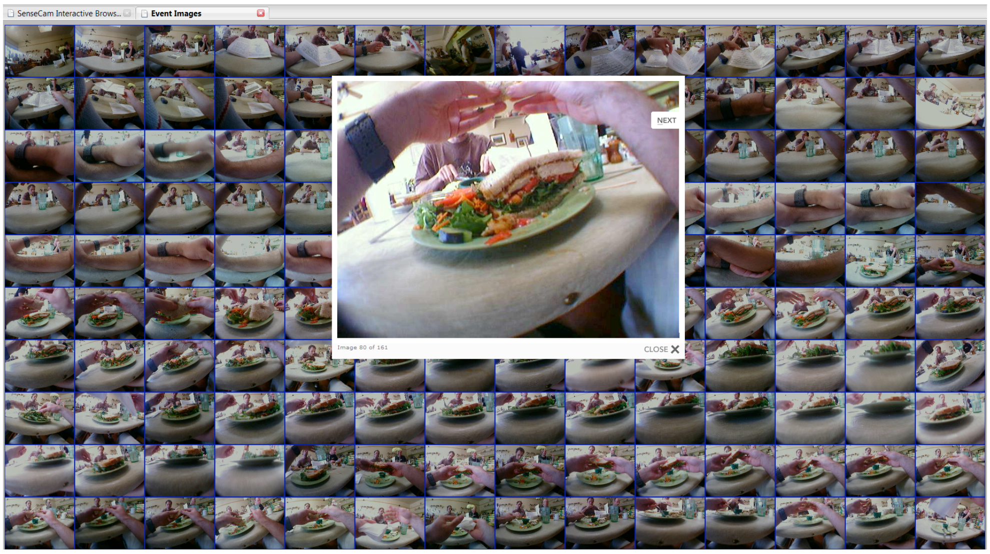


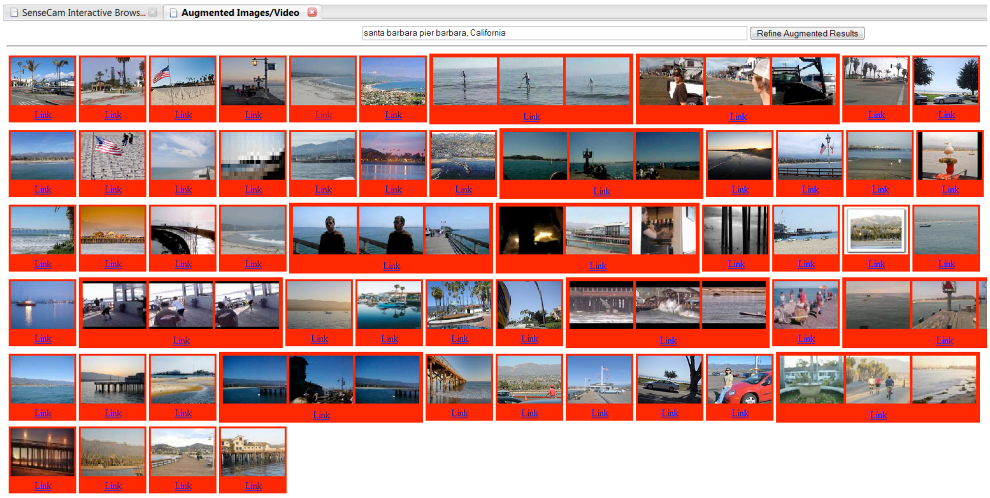
Figure 7.10: Displaying all event's images in new tab on the advanced browser.

in its original size. If the user is dissatisfied with the retrieved results, they are provided with the option of amending the intelligently constructed query in the text box at the top of the screen and then clicking on the "Refine Augmented Results" button, after which a new set of augmented image/video results are displayed.

The advanced browser was used by 4 users within the Centre for Digital Video Processing in Dublin City University. These individuals used the system to review their own SenseCam images to relive past days/events. In total the users collected 1,849,220 images over a period of 22 months which were segmented into 21,967 events. The evaluation section will focus on how effectively users' images were segmented into distinct events/activities, how effective the event importance determination was, the usefulness of the selected keyframe, the relevance of other retrieved events to a queried event, and whether augmented images/videos provide additional retrieval cues to users.

7.3 Evaluation

In this section we will now anecdotally review the hypothesis of this dissertation using two lifelog browsers. In the previous 4 chapters all results have been measured in terms of



santa barbara pier barbara, California

Figure 7.11: Displaying other images/videos to augment the SenseCam event.

precision, recall, F1-Measure, mean average precision, etc. These metrics are very precise, but in this chapter we are interested in discovering how the various aspects work in reality in a lifelog management system. Therefore all of the results in this section are concentrated on feedback offered by the users of the various systems. We will now discuss our findings on the usefulness of each major facet after discussions with the users.

7.3.1 Event Segmentation

We feel that it is useful to segment sequences of images into distinct events or activities. The premise of this is that *"... segmenting ongoing activity into events is important for later memory of those activities ... "* [158]. In Chapter 3 of this thesis we have detailed that in terms of the F1-Measure our optimal event segmentation technique is over 60% accurate against a semantic groundtruth. However, in this section we now investigate how useful event segmentation is to end users in two experimental setups, one the "event-based" browser and the other being the "advanced" browser.

Users found the "event based" browser to be very easy to use, given that it segmented images into events, and had no additional functionality that may have confused any potential users. Our user in Utrecht, the Netherlands, commented that *" ... initially I was not sure if* *a pure time-based segmentation would have been sufficient already (e.g. new segment every 30 min). However, after using it, I think the one provided by your system is much better . . . In Cape Town, I made a tour to the Cape of Good Hope. When browsing the pictures, I was interested in finding the ones that actually show the Cape. I remembered which day it was and given the segmentation, it was easy to find the related segments . . . "* Our user in Utrecht did find a certain example where the event boundaries were undersegmented, i.e. no boundary was identified between his activities of *"sitting on a table eating a sandwich in the hotel's restaurant"* and *"walking around in the lobby checking out some flyers with tourist attractions"*. However, on the whole our user in Utrecht concluded that *"Overall my bottom line is that the segmentation is very useful, even if it is not perfect all the time. It helps you in browsing and finding information."*

The users in Tampere were not so concerned regarding the precision of the event segmentation, instead they were very focused on the event segmentation having a high recall i.e. they don't mind if many events are incorrectly segmented by the system, however they do want to make sure that *all* semantic boundaries are detected though. In their work they used the SenseCam as a tool to find out more about the normal working practices of molecular medical researchers. As a result they wanted to find out about every small event/activity that the medical researcher was involved in. Therefore the information specialists did not mind quickly sifting through the keyframe images of many events, being comfortable in the knowledge that they would identify *every* activity the user was involved in. To accommodate the wishes of these users, we were able to lower a thresholding parameter so as to present more segmented events.

The memory scientists in Toronto found the event segmentation aspect of the "event based" browser to be *". . . very useful and is capable of identifying events with a high degree of accuracy . . . "* One of the researchers commented that *". . . There was rarely a time when an event was not identified . . . "*, and indeed the other researcher commented that *". . . the calendar feature allows for easy recall of the events experienced in different days and gives an overall theme for each day. On certain events, I felt like the computer actually* *understood the photos and their sequence, segmenting them into appropriate events with sharp boundaries separating them. . . . ".* While pointing out some examples where the events were over-segmented, in general the researchers in Toronto found our process of segmenting images into events to be highly effective.

While only receiving the "event based" browser recently, the memory scientist in Duke University commented that *". . . it seems the software is effective . . . ".* He noticed that *". . . the segmentation was particularly good during times when the person went from an inside location to an outside location . . . "*, while acknowledging that the segmentation process sometimes wrongly *". . . created several "events" while that person was driving in her car . . . ".* However, in general the memory scientist in Duke University was happy with the effectiveness and usefulness of the event segmentation facet in the hypothesis of this dissertation.

Users of the "advanced" browser found that the event segmentation was overall very useful. User number 2 notes that the segmentation is *". . . very good as it give me an instant summary of my day . . . ".* Meanwhile user number 3 commented that segmentation is *". . . helpful in general as it breaks all of my images up into easier to understand segments . . . ".* User number 4 commented that *". . . overall the segmentation is good as it helps me see the sequence of my day from breakfast, to giving a presentation, to visiting people in a studio . . . ".*

While the event segmentation is an important and useful facet, users were able to identify certain instances where the segmentation algorithm incorrectly identified event boundaries e.g. user 4 noted that *". . . my presentation is incorrectly over-segmented into 3 events . . . .* The reason for this particular over-segmentation is that the user made two very sudden movements during his presentation, and these movements were sufficiently high so as to be above the event segmentation threshold value. User 3 noted that *". . . the boundaries don't always match my semantic interpretation of what an event is . . . "* however this user didn't identify any specific circumstances with which he was unhappy. User 2 meanwhile was very happy with one particular boundary noting that *". . . it has correctly identified half time* *when I was at a football match! . . . ”*.

7.3.2 Event Importance

Given that a user will still have over 20 events in an average day we argue that it is useful to emphasize more interesting events. The premise is that users aren’t overburdened in terms of having to sift through a number of routine/mundane events before finding a more interesting event which is quite likely to be strongly encoded in memory. Instead more interesting events will appear larger to the user on screen, as described in the experimental setup of the “advanced” browser. In Chapter 5 of this thesis we have detailed that our users rated the event importance algorithm as good. Now we investigate how this works in practice, and also how useful our subjects found this feature to be.

Overall users found that emphasising more important/interesting events was very helpful. User 2 noted that the event importance was either *“excellent”, “very good”* or *“superb”* for various randomly browsed days. In fact overall user 2 noted that *“. . . I like events being highlighted as they gave me a focus or trail to start browsing with . . . ”* User 4 noted that he was *“. . . happy that the mundane events of working on my computer are either very small or hidden. I find the top 2 most important events proposed by the system pleasurable to review, and all the other events aren’t interesting to look at, so the system is correct . . . ”*. However, user 3 felt there was little benefit to the event importance as he commented on being unable to trust the system to consistently identify the most important events; but against this he still felt it is a useful feature and noted that *“. . . I like the fact that events I manually marked as ‘favourites’ are then highlighted as important events . . . ”*

While all users were agreed on the usefulness of highlighting important events, given these comments by user 3, it is evident that not everyone was completely satisfied with the actual performance of the algorithm. However, the majority of users found the events suggested as important/interesting/unique/distinct to be very effective. User 2 noted for a particular day that *“. . . the event importance is excellent as it had the Times Square and Ground Zero events as big, and also the event of talking to the interesting lady from Co.*

Down in Battery Park too. My only complaint is that in the evening I was chatting to an old lady in the diner but it was not recognised as important. . . . ". An explanation for why this particular event of chatting to the old lady was not automatically determined as important, is that the user took off his SenseCam while eating and then focused it on his meal rather than the person he was talking to. As the images were of the wearer's meal and thus by not being able to detect faces, the algorithm could not determine the event as being very important. User 4 was also happy with the automatic determination of event importance too, but had one specific complaint, namely *". . . the presentation I gave on the 2nd of May should have been identified as my most important event . . . "* An explanation for why this event was not determined as the most important may be due to the fact that people attending the conference were too far away from the speaker to have their faces automatically detected, and hence the algorithm could not determine the event as being very important.

Interestingly User 1 noted that while the most important events presented by the system were *"quite good"*, he noted that some of his events of social interaction reoccurred quite regularly, e.g. having coffee with a work colleague. User 1 felt that these events, even though they contain a high element of social interaction, should not be weighted as strongly due to the fact that they are not highly novel.

7.3.3 Keyframe Selection

It has been established that people feel that it is useful for a lifelog of images to be segmented into events, and also that it is quite useful to emphasise those events that are more important or interesting to review. However, events consist of almost 100 images on average, therefore we feel that it is quite challenging to select a single representative image from each event for visual display. In Chapter 5 of this thesis we have shown that our users generally rated the keyframe selection algorithm as good, and we now explore how this works in practice.

Overall users made very few remarks on this facet of the system. When prompted for a comment, users made very general comments such as the keyframes selected are *"good"*. In fact when user 3 was asked to make a more detailed comment, he responded that *"... the performance overall is good, when I see each keyframe I know what the event is about, however given the fact that they're my events I have a good knowledge what the event is about regardless of the image given ... "*. However, for one particular event user 2 didn't recognise what the event was about by looking at the keyframe image only, *"... I had to play through some of the images of my event of leaving the football stadium before I recognised what the event was. The keyframe was badly chosen as it showed the exit stairway which I did not recognise at all ... "*.

Overall users did not seem very enthusiastic about this aspect of the system, only commenting on the few cases where it incorrectly selected a representative keyframe image. This perhaps suggests that this facet of the system, while not having a "wow" factor, is quite important in terms of not being wrong, as otherwise users will be left frustrated. Therefore a good keyframe selection technique is important for users, and our algorithm works quite effectively.

7.3.4 Retrieving Similar Events

On average each user will log approximately 7,000 events per year. Given that the human mind stores items associatively, we feel it will be vital to provide facilities to users to browse through their lifelog collection to find other potentially similar events to an event of interest. Otherwise users will be overwhelmed with an excessive amount of data to sift through. As described in the experimental setup section of this chapter, the "advanced" browser provides users with the option to "find similar" events to any given event. In Chapter 4 of this dissertation we have carried out extensive experimentation investigating the effectiveness in terms of precision, recall, MAP, etc. However, we now seek to identify the potential usefulness of such a facet in a lifelogging system.

All users found the retrieval facility very useful in the "advanced browser". User 2 noted that *"... sifting through the query results I am reminded of many other interesting* *events and days that I want to review . . . ".* User 4 also found the retrieval facet of the system *"very useful"*.

However while this feature is very useful, all users also highlighted the need for better accuracy. This mirrors the findings of Chapter 4 where we found that the retrieval performance of some topics was excellent, but for other topics the performance was quite poor indeed. User 2 had a query where he was walking along a footpath beside a shorefront, and wished to see other times he was at a waterfront, however the user commented that *". . . the system brings back more 'footpath' events, however this does remind me of other interesting events that I'd like to look at . . . "*. User 3 was satisfied with his results for query of walking in the centre of the city he was living in, commenting that *". . . the first rank was wrong, but all the other results returned on the first page were relevant and also interesting . . . "*. User 4 commented that *". . . when trying to find other presentations I gave, I could only find one other relevant event. The search facility is useful, but needs better accuracy . . . "*.

User 1 had a desire to search for events by time and location, in a fashion quite similar to that carried out in the domain of organising traditional photographs e.g. O'Hare *et. al.* [114]. These features were not integrated into the application presented to the users in this experiment, but could be integrated into future lifelogging systems. User 1 commented that *" . . . I don't want to search only by giving an event as an example, I also want the ability to search by time and/or location . . . "*. However, this user was satisfied with the actual results returned for each query he requested, as the results were amusing and set the user off on different browsing trails which he enjoyed very much.

7.3.5 Event Augmentation

Knowing that lifelogging devices such as the SenseCam capture low-quality images, we believe that it will be useful in the case of very interesting events to provide additional memory cues through augmenting those low-quality images from the wearable camera with higher quality images from external data sources. An example of this would be after attending a football match in Wembley where the user will have the option to augment those images with location-tagged images from an external source like the Flickr database (currently over 65 million geotagged images and owing to its content being driven by a very large community of users, this is expanding by approximately 1 million images every fortnight). Another example may be of a user wearing their SenseCam on holidays to Paris, and wishing to see pictures taken by others of the Eiffel Tower for example. In Chapter 6 we investigated the effectiveness of augmenting lifelog events in terms of precision, recall, etc. but now we seek to address how beneficial this actually is in providing additional memory retrieval cues to users.

There was a disparity in opinion on the usefulness of augmenting lifelog events. User 4 was quite negative about this aspect of the system stating that *"...I don't like that the results come up in a new window, this confuses me as it doesn't feel sufficiently integrated ... "*. It must be stressed however that this comment is more focused on an interface and software development issue as to how the feature should be integrated into the system, rather than being focused on the dissertation facet of event augmentation itself. User 2 was very positive about the benefits offered by event augmentation stating that *"...I like it as it reminded me of recent news of a crane crash in New York that I'd forgotten about, and I'd also forgotten that it happened near where I was staying. It is fun seeing these augmented images as they make me think about other things related to my time there as a tourist ... "*. User 1 was also very supportive of the event augmentation usefulness and commented that *"...all the pictures of my trip to the Cliffs of Moher in Ireland are relevant and add to my enjoyment of reviewing this event! ... "*. Due to the superior retrieval performance when allowing users refine query text, as detailed in Chapter 7, we decided to allow users refine their augmented results in our system. User 4 did not like this feature at all, stating that *"...the text box is a killer, seeing other people's photos feels like a diversion to me, so at most I only want to see one level of iteration, refining the search terms just confuses me ... "*. However, user 2 stated that *"...even though the original automatic query was quite a good approximation, I like refining the search and the fact that results from numerous sources are available in one screen ... "*

A possible explanation on why user 4 didn't find this facet of the system so useful was due to the fact that he struggled to recognise many of the proposed augmented images. This user went on 2 work related trips, and thus didn't have much time to soak in the various tourist sites of the locations he visited. He didn't recognise any of the images returned of his time in Amsterdam city centre, or of the main area in the University of Limerick; an explanation for not recognising these is that his own lifelog events in these locations are short events. While on the other hand user 2 augmented events of free weekends abroad and thus recognised many of the images, commenting that *"... the results for the trip to ground zero in New York are excellent as I recognise many of the images, plus it was also good to see pictures of the world trade towers before they collapsed ... the augmented images for my trip to the ACM Multimedia conference in Santa Barbara are brilliant, as the system automatically determined the name of the conference, and also returned relevant results ...* " Likewise User 1 recognised the augmented images and videos presented to him, but appeared to most enjoy the fun and novelty of new material being presented to him. Indeed user 1 pointed towards future research that could be carried out on the interface challenges in displaying the augmented images in a fun manner to help better visualise cues for memory recall.

7.4 Summary

In previous chapters we have carried out extensive evaluation on various facets of managing a lifelog: segmenting sequences of images into distinct events/activities, determining the importance of events, selecting a keyframe image that is representative of a given event, searching for those events that are similar to given events, and finally augmenting the low-quality images of lifelog events with images from external sources of information. However, the evaluation of those aspects has been carried out in terms of how effective they are in terms of precision, recall, mean average precision, etc. The focus of this particular chapter is to evaluate the hypothesis through carrying out an experiment to investigate how *useful* each of these functions are in exploiting aspects of the human memory system in providing effective retrieval cues to a small group of real users of a SenseCam browsing system.

We built two systems, the "event-based browser" and the "advanced browser" to evaluate the usefulness of various facets of the system. The "event-based browser" was used by 5 users in total including one user in Utrecht (the Netherlands) to browse his own images, 2 memory scientists in Toronto (Canada), 1 memory scientist in Duke University (USA), while it was also used by information researchers in Tampere (Finland) to observe the daily activities of 6 medical researchers. The "event-based browser" was only used to investigate the performance of segmenting sequences of images into distinct events/activities. The "advanced browser" was used to investigate the usefulness of all our proposed facets, and this browser was used by 4 people within DCU to review *their own personal* SenseCam images. Many of our users had good fun in browsing through their lifelog collection on the "advanced browser". User 1, who had a SenseCam collection in excess of 2 years in duration, enjoyed the system particularly well, commenting that " *. . . I loved looking back at pictures of Christmas of 2 years ago, this really brought back many great memories . . .* ".

In our evaluation carried out in this chapter we have identified that all users found event segmentation to be highly useful. This aspect works quite well, but there were certain cases identified where the algorithm performed incorrectly.

Many of a user's daily events are quite routine/mundane, and thus not very interesting to review. Knowing that the human mind strongly encodes distinctive memories we feel that it is useful to automatically emphasise those events that are more interesting, important, or unique. In this chapter we have identified that most users found it useful to have more important events highlighted to them. Users like the fact that manually marked events are also displayed as being very important too, thus giving them a sense of control.

Given that it is useful to segment sequences of images into events to help a user review a day, we feel that it is important to accurately select a keyframe image from each event to be displayed on the lifelog browser. In this chapter we have found that users don't especially notice the keyframe images, often stating that they have a semantic knowledge of what the event is about anyways. In fact the only comments made on the keyframe images were at times when they were poorly selected, thus indicating that users *just expect it to work*. Overall users were happy with the selected keyframes.

Even by segmenting sequences of images into distinct events or activities, there will still be 7,000 events recorded each year on average per user. As a result we believe that it will be helpful to provide users with efficient retrieval functionality to identify relevant events to any given event of interest. All users recognised that retrieval is extremely useful, however they also recognised that it is quite difficult to have good performance across all types of queries.

Finally given that users sporadically attend special events (e.g. a tourist trip to the pyramids of Egypt, or to the soccer world cup final, etc.) we feel that a useful and fun feature to include is the ability to augment lifelog events with images/videos captured by many other users. This augmented material is captured from the Internet. In our experiments on the "advanced browser" we have found quite polarised opinion on the usefulness of this facet of the system. One user almost detested this aspect of the system, while two others found it positively amusing. This feature will have to be better integrated into the lifelog browser in future.

Using SenseCam images we are now able to automatically provide effective memory retrieval cues while also offering users an enjoyable browsing experience. Considering these findings we believe that the hypothesis of this dissertation has been proven.

7.5 Conclusions and Contributions

- Our "event-based" browser has now been distributed to approximately 10 research institutions across the world. This browser is available on request. At the time of writing, our "advanced" browser is being prepared to be made available to other research institutions too.

- Event segmentation, event importance, and event retrieval were all anecdotally found to be very helpful. Our users gave mixed feedback on the usefulness of event augmentation, and in future it must be explored how this functionality should be presented to users. Finally users didn't pass significant comments on the accuracy of the keyframe selection process.

- While event retrieval is a very useful function, we feel that in future it will be interesting to provide "free text" or multi-faceted retrieval functionality, where users can easily search not only via "query-by-example" but also for events in a certain location, or for events of a certain temperature condition, or for events taken on a certain day of the week, etc. It would be a worthwhile future area of research to work with many users to determine how this functionality should be most effectively offered.

Chapter 8

CONCLUSIONS AND FUTURE WORK

In the introduction to this thesis we have pointed towards literature stating that the significant events in our lives define our being, and it is important to remember these events. In fact autobiographical memory affects us to the very core and defines who we are. In the past exaggeration and repeated storytelling were used to remember, and in fact shared memories are a part of our social self, and define who we are as groups of people. Given that populations are getting older and that those developing memory impairments are ever increasing [9], scientists have begun exploring the use of technology to aid memory, through providing strong cues to help individuals recall episodes from their autobiographical memories.

Lifelogging is the term used to describe recording different aspects of your daily life, in digital form, for your own exclusive personal use. It is a form of reverse surveillance, sometimes termed *sous*veillance, referring to us, the subjects, doing the watching of ourselves. Lifelogging can take many forms, such as the application which runs on your mobile phone to 'log' all your phone activities and then present all those activities in a calendar format.

The goal of *lifelogging* is to move individuals towards total memory/experience recall. The lifelogging community attempt to achieve this goal by automatically capturing electronic data from numerous sources of information, e.g. web pages visited, e-mails sent and received, audio recordings of conversations, etc. As image information is strongly encoded in the brain an interesting area in this field of research is visual lifelogging, i.e. an individual capturing *their* activities through the medium of images or video. It is important to do so through passive capture, meaning the use of devices that automatically take images or shoot video, thus requiring no conscious effort by the user to take images which leads to him or her acting in a more natural manner.

In **Chapter 1** we introduced the SenseCam which is a small, wearable camera developed by Microsoft Research in Cambridge, UK that creates a visual record of the wearer's day [71]. The SenseCam is worn on the front of the body, suspended from around the neck with a lanyard[1]. In fact as noted in **Chapter 2** the SenseCam has become the prevalent lifelogging device to be used in studies on the benefits of personal visual diaries to people with neurodegenerative conditions such as Alzheimer's. Therefore it is important to work within the constraints of this device (e.g. low quality of images captured) to give the work of this thesis a greater impact.

In **Chapter 2** we provided an overview on the human memory system and how technology may be of assistance to it on occasion, followed by providing a history of the field of lifelogging. The human memory system is composed of 3 main memory stores: sensory memory, short-term memory, and long-term memory. We have learned that the short-term memory store has a limited capacity, although this capacity can be somewhat increased through the process of "chunking". We detailed that long-term memory can be subdivided into procedural memory (e.g. remembering skills) and declarative (which is further subdivided into semantic and episodic memories). Semantic memory refers to knowing facts (e.g. Paris is capital of France), whereas episodic or autobiographical memories refer to remembering specific events about ourselves. We have learned that people can better remember autobiographical memories when provided with good cues, and we have detailed that visual images are strongly encoded in the brain. However, human memory is far from perfect and throughout history we have devised many techniques to help us remember what

[1] http://research.microsoft.com/sendev/projects/sensecam/

we did. Traditionally the written diary has been a very popular means for people to record autobiographical events that have happened in their lives. Vannevar Bush recognised the potential benefits of automatically managing the overwhelming amount of information and events that we encounter [28]. In 1945 he envisioned a "MeMex" device which would extend the human memory, and which would be a *'... device in which an individual stores all his books, records, and communications, and which is mechanized so that it may be consulted with exceeding speed and flexibility ...'*. We also noted that in the past the visual lifelogging community have concentrated on the challenges of miniaturising the hardware devices, and of how to store the *vast* quantities of data. Only recently has the focus been placed on how to manage these large quantities of data so as to provide effective memory retrieval cues.

In **Chapter 1** we provided sample motivation scenarios on how an organised and searchable lifelog of images could be beneficial. One such scenario was of benefit to people who have memory impairments, whereby each day of their lives is recorded using a SenseCam, and thereafter the images are organised into events, with more important/interesting events emphasised to the user, thus offering more effective cues to trigger memory recall. Another scenario was of benefit to tourists where images were organised into events, and thereafter certain events could be augmented with images taken by other tourists of a particular landmark, thus improving the overall experience of their visit.

As concluded in **Chapter 1**, a SenseCam will capture hundreds of thousands of images per year. The SenseCam has become a widely used device not only in the lifelogging domain, but also in the domain of memory studies. However, there exists a substantial challenge in managing the large collection of personal information collected by this device, which is needed to provide strong memory retrieval cues. To realise this, images will have to be segmented into distinct events e.g. having breakfast, being in work, at lunch, etc. The next challenge will be to automatically determine which of those events are most important or unusual e.g. going to a football match will be more memorable than working in front of a computer. Thereafter it will be necessary to have the ability to search for similar events to an example event e.g. *'find me other times when I was talking to my friend John'*. The final challenge will be that of augmenting the low-quality images from the wearable camera with higher quality images from external data sources, e.g. after attending a football match in Wembley the user will have the option to augment those images with location-tagged images from an external source like the Flickr database.

Therefore the hypothesis of this dissertation is that:

> *In order for visual lifelogging technologies to effectively provide autobiographical memory retrieval cues it is necessary to automatically:*
>
> - *effectively segment a large collection of personal images into distinct events*
> - *accurately determine which of those events are most unusual*
> - *usefully provide the ability to link similar events together*
> - *augment images from the low-resolution wearable device with higher quality images from external data sources.*

8.1 Conclusions

In this section we now provide conclusions on our findings in relation to each point of the stated hypothesis in this dissertation, before reviewing the hypothesis as a whole.

8.1.1 Event Segmentation

> ... *effectively segment a large collection of personal images into distinct events* ...

Given that SenseCam users typically capture approximately 2,000 images each day, we have argued that it is necessary to segment sequences of these images into distinct activities or events. Therefore rather than showing the user so many images, it is better to chunk the images semantically together and display 20 events/items. In **Chapter 3** of this dissertation we investigated a number of techniques to effectively realise this goal of segmenting sequences of images into events.

We have proposed the introduction of *TextTiling, peak scoring*, and *mean thresholding* to the challenge of segmenting sequences of lifelog images into distinct events or activities. Instead of comparing adjacent sensor readings to get a "difference" score, our *TextTiling* technique compares blocks/groups of values thus nullifying the effect of outlier values. Our *peak scoring* technique is a method that automatically emphasises time instances where potential event boundaries may occur. Finally the *mean thresholding* technique is a method that allows training to determine the optimum threshold value on whether a "difference" score is sufficiently high to indicate a potential event boundary or not. In experiments to evaluate the effectiveness of our proposed techniques, 5 users each wore a SenseCam for one month, collecting a total of 271,163 images. These users then created a manual groundtruth of 2,986 boundaries between all the activities.

Through the introduction of our proposed techniques, and also through the fusion of content and contextual sources of information, a significant improvement is achieved over previous state of the art work in this young area of research (29.2% better than Yeung & Yeo [157], 41.4% better than an earlier publication of our own [50], and 90% better than Wang *et. al.* [155]). Using sensor sources only, event segmentation processing is instant and performs with an overall F1-Measure of 0.6072.

8.1.2 Finding Similar Events

. . . usefully provide the ability to link similar events together . . .

Even by segmenting SenseCam images into distinct events or activities, SenseCam users will still typically capture approximately 7,000 events per year. This would present quite an intensive browsing experience, thus in **Chapter 4** we have determined that it is necessary to investigate methods to automatically find other similar events to a given SenseCam event. Indeed as motivated earlier, the human mind stores episodes in an associative manner, thus when a user looks at a certain semantic event, it is quite likely that their mind drifts to memories of other semantically similar or associated events. Therefore we aim to provide this facility to search for other associated events to any given event in a lifelog.

This thesis notes that content and contextual sources should be combined in attempting to retrieve similar events to a given lifelog event. Past work in the lifelogging domain has concentrated on either content alone [89, 53, 155] or context alone [146, 31]. We show that the combination of image content and contextual on-board sensor values is 52.6% better than the use of context only sources, illustrating the importance of using content derived from the images. The use of intelligent segmentation techniques can also lead to a 33.6% retrieval improvement in terms of MAP score, thus indicating the importance of the event segmentation work carried out in Chapter 2.

To retrieve other events similar to a given SenseCam event in a lifelog it is necessary to firstly determine *how to represent* SenseCam events, and then *how to compare* those event representations against each other. We investigated a number of techniques in the very important *representation* stage. We set up two datasets to consider a broad range of user information needs, one containing 40 (24 for training, 16 for testing) "general" queries/topics (e.g. driving, eating, etc.), and another containing 23 more "specific" queries/topics (e.g. events of talking to Joe, visits to the museum, etc.).

We found that the approach towards modelling the event is highly influential in terms of retrieval performance. Instead or representing an event by a single keyframe image, if we represent an event by taking the average feature vector value of all the images present in the event, this leads to a 38% increase in retrieval performance on the set of "specific" queries, and an impressive increase of 106% on the set of "general" queries. However, as discovered in Chapter 4, by only processing the middle 35 images of each event (just over 30% of the entire set of images), 90% of the retrieval performance of considering all the event's images is achieved.

While performance is quite good with respect to the "general" queries (e.g. P@5 for *event average* approach = 0.69 overall), it is challenging to retrieve relevant results for queries seeking a very "specific" information need (where P@5 for *event average* approach = 0.30 overall). Firstly this is most likely due to the fact that there are less relevant potential results in the users' lifelog for the "specific" queries. This may mean in instances there could be less than 10 potentially relevant items in the lifelog, if even that, for many of the users' "specific" queries. Another reason for the relatively poor performance of the "specific" queries is that many of them may require an associated semantic meaning, which is very difficult to extract given that by their nature these queries had a very "specific" information need e.g. "show me events when I was talking to Lynda".

Finally we have compared our approach against two distinct content based image retrieval methods: SIFT, & SURF. We have detailed that our approach works slightly better than the results from these two techniques e.g. our approach works particularly well on "beaches" type queries where dominant colours in the scene are important.

8.1.3 Determining the Most Important Events and Their Optimum Keyframes

. . . accurately determine which of those events are most unusual . . .

When a day's worth of SenseCam images is segmented into distinct events or activities, we have argued in **Chapter 5** that it is useful to automatically indicate how interesting each of those events may be. This allows us emphasise more important or interesting events to a user when reviewing their personal visual diary, thus potentially exploiting the fact that the human memory more strongly remembers distinctive events, which may provide even stronger retrieval cues.

The previous state of the art was based on determining conversational events, which contain detected faces, as the most important events. However, we propose integrating face detection with the concept of visual uniqueness or novelty, whereby those events that occur sporadically are more interesting to review than those routine/mundane events.

We evaluated our proposed approach on a dataset of 176,975 SenseCam images collected by 3 users. We discovered that our proposed approach performs *at least* as well as the previous state of the art 80% of the time and performs 4% better overall.

It must be remembered that each lifelog event consists of almost 100 images on average, therefore the selection of a representative, or keyframe, image to provide a user with an instant visual summary of an event will be important in a browsing interface, as argued in **Chapter 5**. Given that up to 40% of lifelog images may be of a relatively low image quality, we proposed a novel technique to select keyframes based on the image within an event that has the best image "quality". After extensive evaluation in Chapter 5 we have found that our *"Image Quality"* approach performs at least as good as the baseline (*"Middle Image"*) on 81.25% of days, and 15.71% better on average overall. One caveat of this approach is that the processing load is greater than simply selecting the middle image.

8.1.4 Augmenting Lifelog Events

> *. . . augment images from the low-resolution wearable device with higher quality images from external data sources . . .*

In **Chapter 6** we have argued that even if we do perfect management of an individual's SenseCam images, it is useful to additionally augment events with images/videos from external sources of information, which could provide additional memory retrieval cues. We also noted the phenomenal expansion of easily-accessible multimedia content due to "Web 2.0", whereby millions of users are uploading billions of images and videos. This dissertation has explored the idea of augmenting one's lifelog experience with images and videos which have been uploaded by hundreds of millions of users on the internet.

Noting that perhaps as little as 2% of the images on the Internet are geotagged, we have developed a technique to use those geotagged images as "seed" images from which we can extract textual information to then have the ability to find *all* potentially relevant images/videos that the user may enjoy reviewing. In our experiments to investigate the effectiveness of this technique, 11 users rated 67 of their augmented SenseCam events. We have found that our technique in constructing a new textual query provided users with approximately 120% extra relevant augmented items. 70% of users found the suggested text quite helpful, and also enjoyed being empowered when allowed to slightly alter the suggested query text.

We believe our approach realises the goal of augmenting passively captured lifelog events with relevant image/videos captured by a large number of individuals, each of whom only had to make a small contribution. We are now entering an era whereby individuals expend little effort in capturing images and making them available, whether it is passively captured lifelog images or a small number of manually captured photos which are uploaded to a media sharing website. However, in aggregating these small contributions over an enormous scale of users we have automatically enriched the experience of individuals reviewing their trip or event, by providing them with a large number of relevant items of information mined from millions of other individuals.

Even though there are research challenges involved in further improving the quality of the lifelog augmentation process, we have now made real one of the important goals of pervasive user-generated content, as an individual's passively captured content in their lifelog is augmented with user-generated content collected by millions of other individuals.

8.1.5 Hypothesis Review

> *In order for visual lifelogging technologies to effectively provide autobiographical memory retrieval cues it is necessary to automatically:*
>
> - *effectively segment a large collection of personal images into distinct events*
> - *accurately determine which of those events are most unusual*
> - *usefully provide the ability to link similar events together*
> - *augment images from the low-resolution wearable device with higher quality images from external data sources.*

In Chapters 3-6 we carried out extensive investigations on optimising individual stages of the stated hypothesis in this dissertation. Finally in **Chapter 7** we recognised that work in the previous chapters of this thesis were focused on evaluating the accuracy (in terms of precision, recall, etc.) of individual facets of the hypothesis in a very systematic manner. However, something as potentially fulfilling as reflecting on one's personal experiences should not only be reduced to precision and recall numbers, therefore we carried out usability tests to explore how fun, insightful, and useful each facet of our work is to end users. This chapter evaluates in an overall manner whether those individual stages actually do exploit lifelogging technologies and provide effective memory retrieval cues.

All 9 of our users commented that the segmentation of sequences of lifelog images into distinct events is a very helpful step in gisting what one did for each day. The feedback from the memory studies researchers is particularly positive regarding this stage towards providing relevant cues so as to more easily retrieve autobiographical memories.

We also evaluated in-house the other 3 strands of the hypothesis. Feedback from 4 users has indicated a strong endorsement for the processes of automatically determining those events that are more important or interesting to review, as well as providing the ability to search for other events. Both of these techniques exploit the fact that the human memory can more easily remember autobiographical events through cued recall than through free recall. Feedback was more mixed on the augmentation strand of the hypothesis, but still we believe that this information does indeed provide the user with additional cues so as to help them retrieve the experiences in the brain of being there with one stating that *"... the augmented images for my trip to the ACM Multimedia conference in Santa Barbara are brilliant, as the system automatically determined the name of the conference, and also returned relevant results ... "*.

In conclusion we feel that our hypothesis has been verified based on the feedback in Chapter 7. The various facets such as event segmentation, determining important events, selecting good keyframe images for the various events, allowing users to browse through associated events, and augmenting events all were very helpful in providing effective memory cues to the users. What is particularly interesting is that memory studies researchers from Duke University and Toronto found our work of particular help, thus providing further evidence that we have manipulated our source of SenseCam images in a productive manner towards providing effective memory retrieval cues.

8.2 Future Work

Given that we have introduced a relatively young area of research in this dissertation, there are naturally a number of different directions of future research that should be undertaken in our opinion.

- **Exploring Effects of Manual Editing of Event Segmentation Output:** As noted in Chapter 3, any application used to segment sequences of SenseCam images into distinct events or activities should be quick (and hence carried out on sensor sources of information rather than on images). We feel that even though the maximum F1-Measure is at just above 60%, this is perhaps the best segmentation performance achievable using only the SenseCam sensor sources of information, particularly given that the performances are matched against a semantic groundtruth. We feel that in future it will be worthwhile to investigate the benefits of providing levels of interaction that allow users to manually edit the event segmentation solutions provided by the system. The challenges associated with this include intelligently modelling/representing the newly formed event so that it can be retrieved in search tasks and also so that it's importance score can be calculated. This challenge is quite considerable as the users will want their newly defined events to be instantly searchable, hence it will not be possible to extract image features from a selection of the middle images in the event; hence motivating the task of introducing intelligent inference techniques. Also a new keyframe must be determined, and again it would take too long (from the user's perspective) to simply select the highest quality keyframe from the middle 35 images of the newly formed event, and as such intelligent inference techniques again must be developed.

- **Investigating Machine Learning Techniques to Recognise Activities:** The work in this dissertation has been focused on detecting activities, while also attempting to determine how important those activities are. However, this dissertation has focused on offering memory cues, and thus has not considered *recognising* what these activities mean in semantic terms. By using SVM[2] facilities, it will then be possible to offer users the ability to search for events through text, i.e. through free recall. Although we have pointed out in Chapter 2 that the human mind responds much better to cued recall (i.e. browsing) than free recall (i.e. searching), O'Hare *et. al.* have detailed how both of these facilities combined together in the image retrieval domain offer an improved experience to the user [114]. In collaboration with colleagues from the University of Amsterdam, we have began exploration work in extracting the semantics of SenseCam events [30], and we believe this area of research in the visual lifelogging domain merits much current and future attention.

- **Exploiting Augmented Images to Construct a Narrative of One's Tourist Trip:** While there remains a significant challenge in augmenting "time-specific" events (e.g. football match, rock concert, etc.), we believe that at least for "place-specific" events (e.g. Eiffel Tower, Big Ben, etc.) it would be interesting to do some form of post-processing on the augmented material so as to construct a narrative of one's tourist trip. Indeed Nack and Hardman have detailed how people like to review multimedia content in a story-like format [105]. Given that augmented images are taken of places where the user has been, we believe that conducting research into generating an amusing/interesting story of these augmented images will add to the overall experience of users, and perhaps make the retrieval cues more effective.

- **Investigating Challenges Posed by Browsing Lifelogs on Mobile Devices:** Given the popularity of mobile/cell phones in the world today, and the prevailing culture of constantly being connected with the world while on the move, we believe that in future people will want to review their lifelogs on handheld mobile devices. Such technology will be helpful to provide memory retrieval cues when one is trying to remember a recent event from their autobiographical memory, while talking to their

[2] Support Vector Machine learning: a commonly used machine learning and pattern recognition algorithm

friend. Not only will a mobile device provide a retrieval cue to the event in memory, it also will allow them to display the images of that event to their friend on the move, thus helping the friend visualise what was happening, and perhaps better understand the emotions involved, based on the visual evidence provided. The typical challenges of mobile devices include dealing with low-bandwidth, a small screen, and as the person is on the move, content must be accessed quickly. In future devices like the SenseCam may allow images to be transferred wirelessly (e.g. via Bluetooth [71]) to mobile devices for instant review of events that have occurred recently e.g. reviewing a morning meeting in the afternoon. As a result we believe that a productive area of research will be investigating power efficient algorithms on hardware circuitry so as to organise lifelog images into events and then determine the importance of those events (based on social interaction i.e. detecting faces) on mobile devices. It is interesting to note that such work on mobile devices is already being investigated for other multimedia processing algorithms [85].

- **Investigating Challenges in Creating Automated Blogs:** While we feel that the primary focus of the SenseCam is as a potential memory aid to those who need it most (those with neurodegenerative conditions), we do believe that others may be interested quite in the fun aspects of maintain a visual lifelog. Indeed O'Hara, Tuffield & Shadbolt believe that it is quite likely that lifelogging and social networking activities will intersect, *"…some social networkers will use lifelogging techniques to generate large quantities of information for their own use, and who will not be shy about sharing it with friends or like-minded people … "* [113]. We believe that in the future it will be interesting to investigate the challenges in automatically creating a visual blog or social network entry for each day that an individual wears a SenseCam or other lifelogging device. It is quite likely that challenges would still remain in trying to identify the distinct events/activities, and then in determining which of those are most interesting to review. There exists a challenge in optimally presenting a selection of a person's most important events for any given day.

- **Using Visual Lifelogging as an Ethnography Aid:** Ethnography, task-observation and field studies are often carried out in order to better understand the social perspective of task operation in order to design systems, which more appropriately support them. Through observation, systems can incorporate an understanding of the user's work setting, social context and pattern of activities [43]. We believe that the SenseCam may be of great benefit in ethnographic studies due to its passive capture nature [29] and as such believe that more in-depth studies as to the strengths and weakness of this device should be carried out by researchers in this domain. If the SenseCam is found to be promising in this regard (as is the case in [29, 83]), then using the work of this dissertation to find more interesting events, possible applications could include: diet monitoring aids; automated logging of work activities thus reducing paperwork (e.g. police officers); pattern analysis of behavioural activities [30]; etc.

- **Leveraging Other Sources of Information:** There are on-going ambitious lifelogging projects incorporating numerous sources of information, particularly data captured on one's personal computer e.g. e-mail, web pages visited, applications open, etc. [60, 54, 97]. Meanwhile other projects are beginning to explore not only collecting images on the move like the SenseCam, but also GPS (for one's location), Bluetooth (for people possibly in the surrounding area), and biometric information (to detect events that are potentially emotionally significant) information [78]. Indeed we have noted that GPS and Bluetooth information can be useful in retrieving information from one's lifelog [31]. While such technologies are somewhat cumbersome to use at the minute, either in terms of how easy they are to wear, how much battery power they consume, or how much CPU time they take up, it is quite possible that such technologies may be improved and become more acceptable to use in the future. Already in the field of biometrics, sensors embedded in clothing have been developed [138]. Indeed, Anderson & Lee note that convenience of mobile/pervasive devices is more important than any perceived social status benefits [5]. In a study on an iPod jacket they noted that *"... the results indicate that consumers find con-*

venience and compatibility of the product most important . . . ". In addition to the aforementioned sources of information we also recognise that many people have a diverse range of online websites to log various activities, e.g. runners log their training performance using their Nike + Apple technology[3], various people use the Twitter website to make small comments on the activities they are doing throughout the day[4], the phenomenon of social networking sites has been well established (e.g. MySpace[5] has approximately 245 million users), people are now recording their medical health records as well (e.g. Microsoft Health Vault[6] and Google Health[7]), and there are even specialist sites such as ones to record things from womens' menstrual cycles[8] to sites for couples to log their sexual activities[9]! We believe that in future it will be imperative for these sources of information to be aggregated together and leveraged by lifelogging applications to be of greater benefit to individuals.

While there is a risk of being swamped by too much data, we believe that more diverse data enables us to make better decisions on the lifestyles that people are leading. This argument is made by Marissa Mayer, Google's vice president of Search Products & User Experience, who comments that sometimes it's not about the quality of the algorithm, but more so about the quantity, quality, and diversity of the available data which can at least help us identify macro trends [117].

- **Increased Collaboration Across Diverse Backgrounds:** The focus of this dissertation has been on investigating the necessary computational techniques to be carried out so as to utilise visual lifelogs to provide users with relevant cues so as to aid memory retrieval. We believe that this is a very important area of consideration and that the hypothesis of this dissertation has been validated. However, in moving forward it will be necessary to determine the information need of users with neurodegenerative

[3] www.nikeplus.com
[4] www.twitter.com
[5] www.myspace.com
[6] www.healthvault.com
[7] www.google.com/health
[8] www.mymonthlycycles.com
[9] www.bedposted.com

conditions, and indeed it will be necessary to determine if the cues provided by techniques such as those in this dissertation are of real benefit in studies on these subjects. As our background is in information processing and management, it will be necessary to forge stronger collaborations with research groups specialising in memory studies. As noted in Chapter 7, we have begun forging such collaborations through releasing software to various institutions in Europe and North America. Such collaborations will be made much stronger in the future, which should result in a more rounded and enlightened approach towards creating the right technology to be of most potential benefit to those who have memory impairments.

However, as has become evident in this section, the field of lifelogging will have a large number of other potential applications. It is our firm belief that personalised health applications are an important area of research currently and especially in the future. We have mentioned that there are many sources of data that can be leveraged, and particularly SenseCam data, as it will help build up an overview of one's lifestyle [30]. While we, as information processing scientists, have the expertise to address the significant research challenges of automatically leveraging streams of lifelogging data so as to make inferences on one's lifestyle, it will be necessary to have strong collaboration with medical, health, fitness, and human performance scientists. Again we feel that it will only be through actively pursuing such collaborations that it will be possible for the scientific community as a whole to exploit the vast potential of lifelogging technologies for everyone. O'Hara comments that *"... To that extent, lifelogging tools are tools for everyone to exert more control over their personal data, their public presence online and their digital identity ... "*, and through proper collaboration with experts in other domains with an intimate understanding of the actual human requirements for such technologies, the gains for the human race as a whole will be much greater.

8.3 Summary

At the start of this thesis we stated that the capability to retrieve autobiographical memories from our brain defines our very self. We have shown that there are a number of memory systems in the brain: sensory memory, short-term memory, and long-term memory. We are interested in the autobiographical aspect of the long-term memory system, however we have also exploited aspects of the short-term memory system too. Given that there are a large number of people with conditions that affect autobiographical memory, scientists have recently begun to investigate the benefits of using technology to somewhat help alleviate this problem. The field of lifelogging has been attempting to provide thorough answers to this difficult problem, and we have given an overview of the work carried out thus far. However, given that visual images provide strong cues to help people retrieve autobiographical memories, the visual lifelogging field has received most attention. Most efforts in this domain have been concentrated on hardware miniaturisation, and on storage problems, but these problems are almost solved with the SenseCam now being the prevalent device in this domain. Given that human memory responds much better to cued recall than free recall, and also that images taken from the viewpoint of where an autobiographical memory was encoded provide strong retrieval cues, the SenseCam has displayed much promise as a memory aid. However, given the volume of data generated by this device, we have hypothesised that a number of steps have to be taken for the SenseCam to truly become an effective memory aid.

In this thesis we have made contributions in techniques to segment sequences of images into distinct events thus making it easier for users to digest large volumes of SenseCam images. The selection of a representative image from an event, so as to provide a good memory cue to a user as to what the event is about, is an important challenge and we have introduced a new technique to incorporate the notion of image quality. Thirdly we have proposed investigating the novelty of given events to determine how interesting they may be to review, exploiting the fact that distinctive memories are strongly encoded in the mind. Also given the fact that the brain stores memories in an associative manner, we have proposed techniques to effectively allow users to search for potentially similar events to a given SenseCam event. Finally we have introduced the novel idea of augmenting one's lifelog collection with material from other peoples' images/video which are publicly available on the Internet. In utilising this augmented material we provide further retrieval cues to help people remember their given events.

Based on feedback from our users in Dublin City University and elsewhere, including researchers in Duke University and in Toronto who are investigating the benefits of the SenseCam to human memory, we believe that this dissertation is of benefit to the research community. We have now introduced a variety of automated techniques that effectively exploit characteristics of the human memory system so as to provide strong memory retrieval cues from SenseCam images.

Bibliography

[1] Kiyoharu Aizawa. Emerging Issues for Multimedia Analysis and Applications. In *International Workshop on Multimedia Content Analysis and Mining*, volume 4577, pages 14–15, Weihai, China, 2007.

[2] Kiyoharu Aizawa, Tetsuro Hori, Shinya Kawasaki, and Takayuki Ishikawa. Capture and efficient retrieval of life log. In *Proceedings of the Pervasive 2004 Workshop on Memory and Sharing of Experiences*, pages 15–20, Linz/Vienna, Austria, 2004.

[3] Kiyoharu Aizawa, Ken-Ichiro Ishijima, and Makoto Shiina. Summarizing Wearable Video. In *International Conference on Image Processing*, pages 398–401, Thessaloniki, Greece, 2001.

[4] Anita L. Allen. Dredging-up the Past: Lifelogging, Memory and Surveillance. *New Yorker*, pages 38–44, 2007.

[5] Gretchen Anderson and Gwanhoo Lee. Why Consumers (Don't) Adopt Smart Wearable Electronics. *IEEE Pervasive Computing*, 7(3):10–12, 2008.

[6] John R. Anderson, editor. *Language, Memory, and Thought*. Lawrence Erlbaum Associates, 1974.

[7] Daniel Ashbrook, Kent Lyons, and James Clawson. Capturing Experiences Anytime, Anywhere. *IEEE Pervasive Computing*, 5:8–9, April 2006.

[8] R.C. Atkinson and R.M. Shiffrin. Human memory: A proposed system and its control processes. *The psychology of learning and motivation*, 2:89–195, 1968.

[9] Alan Baddeley, editor. *Your Memory: A User's Guide*. Carlton Books, 2004.

[10] Alan Baddeley and Graham Hitch. Working Memory. *The psychology of learning and motivation*, 8:47–89, 1974.

[11] Ron M. Baecker, Elsa Marziali, Sarah Chatland, Kante Easley, Masashi Crete, and Martin Yeung. Multimedia Biographies for Individuals with Alzheimer's Disease and Their Families. In *2nd International Conference on Technology and Aging*, Toronto, Canada, 2007.

[12] Liam J. Bannon. Forgetting as a feature, not a bug: the duality of memory and implications for ubiquitous computing. *CoDesign*, 2(1):3–15, 2006.

[13] Deborah Barreau, Abe Crystal, Jane Greenberg, Anuj Sharma, Michael Conway, John Oberlin, Michael Shoffner, and Stephen Seiberling. Augmenting Memory for Student Learning: Designing a Context-Aware Capture System for Biology Education. *Proceedings of the American Society for Information Science and Technology*, 43:251–251, Oct 2007.

[14] H. Bay, T. Tuytelaars, and L. Van Gool. SURF: Speeded Up Robust Features. In *Proceedings of the 9th European Conference on Computer Vision (ECCV'06)*, pages 404–407, Graz, Austria, May 2006.

[15] Gordon Bell and Jim Gemmell. A Digital Life. *Scientific American*, 2007.

[16] Bettina Berendt and Christoph Hanser. Tags are not Metadata, but "Just More Content" - to Some People. In *Proceedings of the International Conference on Weblogs and Social Media (ICWSM 2007)*, Boulder, Colorado, USA, 2007.

[17] Emma Berry, Narinder Kapur, Lyndsay Williams, Steve Hodges, Peter Watson, Gavin Smyth, James Srinivasan, Reg Smith, Barbara Wilson, and Ken Wood. The use of a wearable camera, SenseCam, as a pictorial diary to improve autobiographi-

cal memory in a patient with limbic encephalitis. *Neuropsychological Rehabilitation*, 17(4):582–601, 2007.

[18] M. Blighe, A. Doherty, A.F. Smeaton, and N. O'Connor. Keyframe Detection in Visual Lifelogs. In *1st International Conference on Pervasive Technologies Related to Assistive Environments (PETRA)*, Athens, Greece, July 2008.

[19] M. Blighe, S. Sav, H. Lee, and N. O'Connor. Mo Músaem Fíorúil: A Web-based Search and Information Service for Museum Visitors. In *International Conference on Image Analysis and Recognition (ICIAR)*, Povoa de Varzim, Portugal, 2008.

[20] Michael Blighe, Herve le Borgne, Noel E. O'Connor, Alan F. Smeaton, and Gareth J.F. Jones. Exploiting Context Information to aid Landmark Detection in SenseCam Images. In *ECHISE 2006 - 2nd International Workshop on Exploiting Context Histories in Smart Environments - Infrastructures and Design, 8th International Conference of Ubiquitous Computing (Ubicomp 2006)*, Orange County, CA, 2006.

[21] Michael Blighe, Sorin Sav, Hyowon Lee, and Noel E. O'Connor. Mo Músaem Fíorúil: A Web-based Search and Information Service for Museum Visitors. In *ICIAR: International Conference on Image Analysis and Recognition*, LNCS, Povoa de Varzim, Portugal, 2008.

[22] Mark Blum, Alex (Sandy) Pentland, and Gerhard Troster. Interest-Based Life Logging. *IEEE Multimedia*, 13(4):40–48, 2006.

[23] Gary Rost Bradski and Adrian Kaehler, editors. *Learning OpenCV*. O'Reilly, 2008.

[24] William F. Brewer. *Practical Aspects of Memory: Current Research and Issues*, chapter Qualitative analysis of the recalls of randomly samples autobiographical events, pages 263–268. Wiley, 1988.

[25] Huw W. Bristow, Chris Baber, James Cross, James F. Knight, and Sandra I. Woolley. Defining and evaluating context for wearable computing. *International Journal of Human Computer Studies*, 60(5-6):798–819, 2004.

[26] Michael Bukhin and Michael DelGaudio. WayMarkr: acquiring perspective through continuous documentation. In *MUM '06: Proceedings of the 5th international conference on Mobile and ubiquitous multimedia*, page 9, Stanford, California, 2006.

[27] Neil Burgess, Suzanna Becker, John A. King, and John O'Keefe. *Episodic Memory*, chapter Memory for events and their spatial context: models and experiments, pages 249–268. Oxford University Press, 2002.

[28] Vannevar Bush. As We May Think. *The Atlantic Monthly*, 176(1):101–108, Jul 1945.

[29] Daragh Byrne, Aiden R. Doherty, Alan F. Smeaton, Gareth J.F. Jones, Sanna Kumpulainen, and Kalervo Järvelin. The SenseCam as a Tool for Task Observation. In *HCI 2008 - 22nd BCS HCI Group Conference*, Liverpool, U.K., 2008.

[30] Daragh Byrne, Aiden R. Doherty, Cees G.M Snoek, Gareth J.F. Jones, and Alan F. Smeaton. Validating the Detection of Everyday Concepts in Visual Lifelogs. In *SAMT 2008 - 3rd International Conference on Semantic and Digital Media Technologies*, pages 15–30, Koblenz, Germany, 2008.

[31] Daragh Byrne, Barry Lavelle, Aiden R. Doherty, Gareth J.F. Jones, and Alan F. Smeaton. Using Bluetooth and GPS Metadata to Measure Event Similarity in SenseCam Images. In *IMAI'07 - 5th International Conference on Intelligent Multimedia and Ambient Intelligence*, pages 1454–1460, Salt Lake City, Utah, USA, 2007.

[32] T. Canli, J.E. Desmond, Z. Zhao, and J.D.E. Gabrieli. Sex Differences in the Neural Basis of Emotional Memories. *Proceedings of the National Academy of Sciences of the United States of America*, 99(16):10789–10794, 2002.

[33] Jaime Carbonell and Jade Goldstein. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *SIGIR '98: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 335–336, Melbourne, Australia, 1998.

[34] Francine Chen, Matthew Cooper, and John Adcock. Video summarization preserving dynamic content. In *TVS '07: Proceedings of the international workshop on TRECVID video summarization*, pages 40–44, Augsburg, Bavaria, Germany, 2007.

[35] S. Cherry. Total recall [life recording software]. *IEEE Spectrum*, 42:24–30, Nov 2005.

[36] Jean-Pierre Chevallet, Joo-Hwee Lim, and Mun-Kew Leong. Object Identification and Retrieval from Efficient Image Matching. Snap2Tell with the STOIC Dataset. *Information Systems Journal*, 43(2):515–530, 2007.

[37] Ciaran Ó Conaire. Dynamic Thresholding Methods. Technical report, Dublin City University School of Engineering, 2006.

[38] Ciaran Ó Conaire, Noel E. O'Connor, Alan F. Smeaton, and Gareth J.F. Jones. Organising a daily Visual Diary Using Multi-Feature Clustering. In *SPIE Electronic Imaging - Multimedia Content Access: Algorithms and Systems (EI121)*, San Jose, CA, 2007.

[39] Martin A. Conway. *Episodic Memory*, chapter Sensory-perceptual episodic memory and its context: autobiographical memory, pages 53–70. Oxford University Press, 2002.

[40] Martin A. Conway. Memory and the self. *Journal of Memory and Language*, 53(4):594–628, 2005.

[41] Matthew Cooper and Jonathan Foote. Discriminative techniques for keyframe selection. In *ICME 2005 - IEEE International Conference on Multimedia and Expo*, pages 502–505, Amsterdam, The Netherlands, 2005.

[42] Costas Cotsaces, Nikos Nikolaidis, and Ioannis Pitas. Video shot detection and condensed representation. A Review. *IEEE Signal Processing Magazine*, 23:28–37, Mar 2006.

[43] Sally Jo Cunningham and Matt Jones. Autoethnography: a tool for practice and education. In *CHINZ '05: Proceedings of the 6th ACM SIGCHI New Zealand chapter's international conference on Computer-human interaction*, pages 1–8, Auckland, New Zealand, 2005.

[44] Mary Czerwinski and Eric Horvitz. An Investigation of Memory for Daily Computing Events. In *16th British HCI Group Annual Conference*, pages 230–245, London, U.K., 2002.

[45] Graham Davies and Donald M. Thomson, editors. *Memory in Context: Context in Memory*. John Wiley & Sons, Chichester, England, 1988.

[46] Aiden R. Doherty, Daragh Byrne, Alan F. Smeaton, Gareth J.F. Jones, and Mark Hughes. Investigating keyframe selection methods in the novel domain of passively captured visual lifelogs. In *CIVR '08: Proceedings of the 2008 international conference on Content-based image and video retrieval*, pages 259–268, Niagara Falls, Canada, 2008.

[47] Aiden R. Doherty, Ciarán Ó Conaire, Michael Blighe, Alan F. Smeaton, and Noel E. O'Connor. Combining image descriptors to effectively retrieve events from visual lifelogs. In *MIR '08: Proceeding of the 1st ACM international conference on Multimedia information retrieval*, pages 10–17, Vancouver, British Columbia, Canada, 2008.

[48] Aiden R. Doherty and Alan F. Smeaton. Automatically Segmenting Lifelog Data Into Events. In *WIAMIS: 9th International Workshop on Image Analysis for Multimedia Interactive Services*, pages 20–23, Klagenfurt, Austria, 2008.

[49] Aiden R. Doherty and Alan F. Smeaton. Combining Face Detection and Novelty to Identify Important Events in a Visual Lifelog. In *CIT: 8th International Conference on Computer and Information Technology, Workshop on Image- and Video-based Pattern Analysis and Applications.*, Sydney, Australia, 2008.

[50] Aiden R. Doherty, Alan F. Smeaton, Keansub Lee, and Daniel P.W. Ellis. Multimodal Segmentation of LifeLog Data. In *RIAO 2007 - Large-Scale Semantic Access to Content (Text, Image, Video and Sound)*, Pittsburg, PA, USA, 2007.

[51] Emilie Dumont and Bernard Merialdo. Split-screen dynamically accelerated video summaries. In *TVS '07: Proceedings of the international workshop on TRECVID video summarization*, pages 55–59, Augsburg, Bavaria, Germany, 2007.

[52] Daniel P.W. Ellis and Keansub Lee. Minimal-impact audio-based personal archives. In *CARPE'04: Proceedings of the the 1st ACM workshop on Continuous archival and retrieval of personal experiences*, pages 39–47, New York, New York, USA, 2004.

[53] Daniel P.W. Ellis and Keansub Lee. Accessing Minimal-Impact Personal Audio Archives. *IEEE Multimedia*, 13(4):30–38, 2006.

[54] David Elsweiler. *Supporting Human Memory in Personal Information Management*. PhD thesis, Department of Computer and Information Sciences, University of Strathclyde, 2007.

[55] Rowanne Fleck and Geraldine Fitzpatrick. Supporting Collaborative Reflection with Passive Image Capture. In *Supplementary proceedings of COOP'06*, pages 41–48, Carry-le-Rouet, France, 2006.

[56] E. Fox and J. Shaw. Combination of Multiple Searches. In *TREC 2 Text REtrieval Conference*, pages 243–252, Gaithersberg, Maryland, USA, 1993.

[57] Georgina Gaughan. *Novelty Detection in Video Retrieval: Finding New News in TV News Stories*. PhD thesis, Dublin City University, 2006.

[58] Georgina Gaughan and Alan F. Smeaton. Finding New News: Novelty Detection in Broadcast News. In Gary Geunbae Lee, Akio Yamada, Helen Meng, and Sung Hyon Myaeng, editors, *AIRS 2005 - Second Asia Information Retrieval Symposium.*, volume 3689 / 2006, pages 583–588, Jeju Island, Korea, 2005.

[59] Jim Gemmell, Aleks Aris, and Roger Lueder. Telling Stories with MyLifeBits. In *ICME - International Conference on Multimedia and Expo*, pages 1536–1539, Amsterdam, The Netherlands, 2005.

[60] Jim Gemmell, Gordon Bell, and Roger Lueder. MyLifeBits: A personal database for everything. *Communications of the ACM*, 49, 2006.

[61] Jim Gemmell, Lyndsay Williams, Ken Wood, Roger Lueder, and Gordon Bell. Passive capture and ensuing issues for a personal lifetime store. In *CARPE'04: Proceedings of the the 1st ACM workshop on Continuous archival and retrieval of personal experiences*, pages 48–55, New York, New York, USA, 2004.

[62] Phillipa Gill, Martin Arlitt, Zongpeng Li, and Anirban Mahanti. YouTube Traffic Characterization: a View from the Edge. In *IMC: Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*, pages 15–28, San Diego, California, USA, 2007.

[63] D.R. Godden and Alan Baddeley. Context-dependent memory in two natural environments: On land and underwater. *British Journal of Psychology*, 66(3):325–331, 1975.

[64] Adrian Graham, Hector Garcia-Molina, Andreas Paepcke, and Terry Winograd. Time as essence for photo browsing through personal digital libraries. In *JCDL '02: Proceedings of the 2nd ACM/IEEE-CS joint conference on Digital libraries*, pages 326–335, Portland, Oregon, USA, 2002.

[65] Jane Greenberg, Abe Crystal, Anuj Sharma, Eva Méndez, John Oberlin, and Michael Shoffner. Memex Metadata (M 2) for Reflective Learning. In *International conference on Dublin Core and Metadata Applications: metadata for knowledge and learning*, pages 169–179, Manzanillo, Colima, Mexico, 2006.

[66] Robert L. Greene, editor. *Human Memory: Paradigms and Paradoxes*. Lawrence Erlbaum Associates, 1992.

[67] Cathal Gurrin, Alan F. Smeaton, Daragh Byrne, Neil O'Hare, Gareth J.F. Jones, and Noel E. O'Connor. An Examination of a Large Visual Lifelog. In *AIRS 2008 - Asia Information Retrieval Symposium*, Harbin, China, 2008.

[68] R Harper, D Randall, N Smyth, C Evans, L Heledd, and R Moore. Thanks for the Memory. In *HCI 2007 - Proceedings of the 21st BCS HCI Group Conference*, Lancaster, U.K., 2007.

[69] R Harper, D Randall, N Smyth, C Evans, L Heledd, and R Moore. The Past is a Different Place: They Do Things Differently There. In *Designing Interactive Systems*, pages 271–280, Cape Town, South Africa, 2008.

[70] Martin A. Hearst and Christian Plaunt. Subtopic structuring for full-length document access. In *SIGIR '93: Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 59–68, Pittsburgh, Pennsylvania, United States, 1993.

[71] Steve Hodges, Lyndsay Williams, Emma Berry, Shahram Izadi, James Srinivasan, Alex Butler, Gavin Smyth, Narinder Kapur, and Ken Wood. SenseCam: A Retro-

spective Memory Aid. In *UbiComp: 8th International Conference on Ubiquitous Computing*, volume 4602 of *LNCS*, pages 177–193, California, USA, 2006.

[72] Tetsuro Hori and Kiyoharu Aizawa. Context-based video retrieval system for the life-log applications. In *MIR '03: Proceedings of the 5th ACM SIGMM international workshop on Multimedia information retrieval*, pages 31–38, Berkeley, California, 2003.

[73] Alexandar Jaffe, Mor Naaman, Tamir Tassa, and Marc Davis. Generating summaries and visualization for large collections of geo-referenced photographs. In *MIR '06: Proceedings of the 8th ACM international workshop on Multimedia information retrieval*, pages 89–98, Santa Barbara, California, USA, 2006.

[74] Daniel Kahneman, Alan B. Krueger, David A. Schkade, Norbert Schwarz, and Arthur A. Stone. A Survey Method for Characterizing Daily Life Experience: The Day Reconstruction Method. *Science*, 306:1776–1780, 2004.

[75] J.N. Kapur, P.K. Sahoo, and A.K.C. Wong. A new method for graylevel picture thresholding using the entropy of the histogram. *Computer vision, graphics, and image processing*, 29(3):273–285, 1985.

[76] Tatsuyuki Kawamura, Tomohiro Fukuhara, Hideaki Takeda, Yasuyuki Kono, and Masatsugu Kidode. Ubiquitous Memories: A Memory Externalization System Using Physical Objects. *Personal and Ubiquitous Computing*, 11(4):287–298, 2007.

[77] Peter Kehoe and Alan F. Smeaton. Using Graphics Processor Units (GPUs) for Automatic Video Structuring. In *WIAMIS - 8th International Workshop on Image Analysis for Multimedia Interactive Services*, pages 18–18, Santorini, Greece, 2007.

[78] Liadh Kelly and Gareth J.F. Jones. Venturing into the Labyrinth: the Information Retrieval Challenge of Human Digital Memories. In *Workshop on Supporting Human Memory with Interactive Systems, at HCI: The 21st British HCI Group Annual Conference*, pages 37–41, Lancaster, U.K., 2007.

[79] Lyndon Kennedy, Mor Naaman, Shane Ahern, Rahul Nair, and Tye Rattenbury. How flickr helps us make sense of the world: context and content in community-contributed media collections. In *MULTIMEDIA '07: Proceedings of the 15th international conference on Multimedia*, pages 631–640, Augsburg, Germany, 2007.

[80] Lyndon S. Kennedy and Mor Naaman. Generating diverse and representative image search results for landmarks. In *WWW '08: Proceeding of the 17th international conference on World Wide Web*, pages 297–306, Beijing, China, 2008.

[81] Ig-Jae Kim, Sang Chul Ahn, Heedong Ko, and HyoungGon Kim. PERSONE: personalized experience recoding and searching on networked environment. In *CARPE '06: Proceedings of the 3rd ACM workshop on Continuous archival and retrival of personal experences*, pages 49–54, Santa Barbara, California, USA, 2006.

[82] Manesh Kokare, B.N. Chatterji, and P.K. Biswas. Comparison of Similarity Metrics for Texture Image Retrieval. In *TENCON 2003: Conference on Convergent Technologies for Asia-Pacific Region*, pages 571–575, Bangalore, India, 2003.

[83] Sanna Kumpulainen, Kalervo Järvelinand, Sami Serola, Aiden R. Doherty, Daragh Byrne, Alan F. Smeaton, and Gareth Jones. Data Collection Methods for Analyzing Task-Based Information Access in Molecular Medicine. In *MobiHealthInf 2009 - 1st International Workshop on Mobilizing Health Information to Support Healthcare-related Knowledge Work*, Porto, Portugal, 2009.

[84] James M. Lampinen, Denise R. Beike, and Douglas A. Behrend. *The Self and Memory*, chapter The Self and Memory: It's about Time, pages 255–262. Psychology Press, 2004.

[85] Daniel Larkin, Andrew Kinane, and Noel E. O'Connor. Towards Hardware Acceleration of Neuroevolution for Multimedia Applications on Mobile Devices. In *ICONIP 2006 - International Conference on Neural Information Processing*, pages 1178–1188, 2006.

[86] Steen F. Larsen, Charles P. Thompson, and Tia Hansen. *Remembering Our Past: Studies in Autobiographical Memory*, chapter Time in Autobiographical Memory, pages 129–156. Cambridge University Press, 1996.

[87] Matthew L. Lee and Anind K. Dey. Providing good memory cues for people with episodic memory impairment. In *Assets '07: Proceedings of the 9th international ACM SIGACCESS conference on Computers and accessibility*, pages 131–138, Tempe, Arizona, USA, 2007.

[88] Joo-Hwee Lim, Yiqun Li, Yilun You, and Jean-Pierre Chevallet. Scene Recognition with Camera Phones for Tourist Information Access. In *ICME - International Conference on Multimedia and Expo*, pages 100–103, Beijing, China, 2007.

[89] Wei-Hao Lin and Alexander Hauptmann. Structuring Continuous Video Recordings of Everyday Life Using Time-Constrained Clustering. In *Multimedia Content Analysis, Management, and Retieval SPIE-IST Electronic Imaging*, volume 6073, pages 111–119, San Jose, California, USA, 2006.

[90] M. Linton. Transformations of memory in everyday life. *Memory observed: Remembering in natural contexts*, pages 117–118, 2000.

[91] Ting Liu, Charles Rosenberg, and Henry A. Rowley. Clustering Billions of Images with Large Scale Nearest Neighbor Search. In *WACV: Workshop on Applications of Computer Vision, 2007.*, pages 28–28, Austin, Texas, USA, 2007.

[92] D. Lowe. Distinctive image features from scale-invariant keypoints. In *International Journal of Computer Vision*, volume 60(2), pages 91–110, 2004.

[93] Eleanor A. Maguire. *Episodic Memory*, chapter Neuroimaging studies of autobiographical event memory, pages 164–180. Oxford University Press, 2002.

[94] Steve Mann. Wearable computing: a first step toward personal imaging. *Computer*, 30:25–32, Feb 1997.

[95] Steve Mann, James Fung, Chris Aimone, Anurag Sehgal, and Daniel Chen. Designing EyeTap Digital Eyeglasses for Continuous Lifelong Capture and Sharing of Personal Experiences. In *Alt. Chi, Proc. CHI 2005*, Portland, Oregon, USA, 2005.

[96] Markos Markou and Sameer Singh. Novelty detection: a review - part 1: statistical approaches. *Signal Process.*, 83(12):2481–2497, 2003.

[97] Catherine C. Marshall and William Jones. Keeping encountered information. *Commun. ACM*, 49(1):66–67, 2006.

[98] Dan P. McAdams. *The Self and Memory*, chapter The Redemptive Self: Narrative Identity in America Today, pages 95–115. Psychology Press, 2004.

[99] Gerard McAtamney and Caroline Parker. An examination of the effects of a wearable display on informal face-to-face communication. In *CHI '06: Proceedings of the SIGCHI conference on Human Factors in computing systems*, pages 45–54, Montréal, Québec, Canada, 2006.

[100] Dean S. Messing, Peter van Beek, and James H. Errico. The MPEG-7 Colour Structure Descriptor: Image Description Using Colour and Local Spatial Information. In *ICIP01 - International Conference on Image Processing*, pages 670–673, Thessaloniki, Greece, 2001.

[101] George A. Miller. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, 63:81–97, 1956.

[102] George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J. Miller. Introduction to WordNet: An On-line Lexical Database. *International Journal of Lexicography*, 3(4):235–244, 2004.

[103] Mark Montague and Javed A. Aslam. Relevance score normalization for metasearch. In *CIKM '01: Proceedings of the tenth international conference on Information and knowledge management*, pages 427–433, Atlanta, Georgia, USA, 2001.

[104] Frank Nack. You Must Remember This. *IEEE Multimedia*, 12:4–7, Jan 2005.

[105] Frank Nack and Linda Hardmann. Generating Multimedia Presentations: It's All in the Game. *ERCIM News, No. 57*, pages 24–25, 2004.

[106] Milind R. Naphade and John R. Smith. On the detection of semantic concepts at TRECVID. In *MULTIMEDIA '04: Proceedings of the 12th annual ACM international conference on Multimedia*, pages 660–667, New York, NY, USA, 2004.

[107] M. Naveh-Benjamin and T.J. Ayres. Digit span, reading rate, and linguistic relativity. *Experimental Psychology*, 38:739–751, 1986.

[108] Darren Newtson, Gretchen Engquist, and Joyce Bois. The Objective Basis of Behavior Units. *Personality and Social Psychology*, 35(12):847–862, 1977.

[109] Bernt Schiele Nicky Kern and Albrecht Schmidt. Recognizing Context for Annotating a Live Life Recording. *Personal and Ubiquitous Computing*, 11(4):287–298, 2007.

[110] Brahim Nini and Mohamed Batouche. Virtualized Real Object Integration and Manipulation in an Augmented Scene. In *CAIP 2005 - 11th International Conference on Computer Analysis of Images and Patterns*, volume 3691 / 2005, pages 248–255, Versailles, France, 2005.

[111] D. Nistér and H. Stewnius. Scalable recognition with a vocabulary tree. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2161–2168, New York, USA, June 2006.

[112] Noel E. O'Connor, Edward Cooke, Herve le Borgne, Michael Blighe, and Tomasz Adamek. The AceToolbox: Low-Level Audiovisual Feature Extraction for Retrieval and Classification. In *2nd IEE European Workshop on the Integration of Knowledge, Semantic and Digital Media Technologies*, pages 55–60, London, U.K., 2005.

[113] Kieron O'Hara, Mischa Tuffield, and Nigel Shadbolt. Lifelogging: Issues of Identity and Privacy with Memories for Life. In *The First International Workshop on Identity and the Information Society*, pages 1–31, Arona, Lago Maggiore, Italy, 2008.

[114] Neil O'Hare, Cathal Gurrin, Gareth J. F. Jones, Hyowon Lee, Noel E. O'Connor, and Alan F. Smeaton. Using text search for personal photo collections with the Medi-Assist system. In *SAC '07: Proceedings of the 2007 ACM symposium on Applied computing*, pages 880–881, Seoul, Korea, 2007.

[115] Neil O'Hare, Cathal Gurrin, Gareth J.F. Jones, and Alan F. Smeaton. Combination of Content Analysis and Context Features for Digital Photograph Retrieval. In *2nd IEE European Workshop on the Integration of Knowledge, Semantic and Digital Media Technologies*, pages 323–328, London, U.K., 2005.

[116] Neil O'Hare, Hyowon Lee, Saman Cooray, Cathal Gurrin, Gareth J.F. Jones, Jovanka Malobabic, Noel E. O'Connor, Alan F. Smeaton, and Bartlomiej Uscilowski. Automatic Text Searching for Personal Photos. In *SAMT 2006 - Proceedings of The First International Conference on Semantics And Digital Media Technology*, pages 43–44, Athens, Greece, 2006.

[117] Juan Carlos Perez. Google wants your phonemes. `http://www.infoworld.com/article/07/10/23/Google-wants-your-phonemes_1.html`, cited September 2008.

[118] Marcus J Pickering, Daniel Heesch, Robert Ó Callaghan, Stefan Ruger, and David Bull. Video Retrieval using Global Features in Keyframes. In *TREC 2002 - Text REtrieval Conference*, Gaithersburg, Maryland, 2002.

[119] M.I. Posner. *Advances in learning and motivation*, volume 3, chapter Abstraction and the process of recognition, pages 44–96. Academic Press, New York, 1969.

[120] Jesse E. Purdy, Michael R. Markham, Bennett L. Schwartz, and William C. Gordon (2nd edition), editors. *Learning and Memory*. Thomson Learning, 2001.

[121] Tye Rattenbury, Nathaniel Good, and Mor Naaman. Towards Automatic Extraction of Event and Place Semantics from Flickr Tags. In *SIGIR: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 103–110, Amsterdam, The Netherlands, 2007.

[122] Sasank Reddy, Andrew Parker, Josh Hyman, Jeff Burke, Deborah Estrin, and Mark Hansen. Image browsing, processing, and clustering for participatory sensing: lessons from a DietSense prototype. In *EmNets '07: Proceedings of the 4th workshop on Embedded networked sensors*, pages 13–17, Cork, Ireland, 2007.

[123] Sid Reich, Les Goldberg, and Stephen Hudek. Deja view camwear model 100. In *CARPE'04: Proceedings of the the 1st ACM workshop on Continuous archival and retrieval of personal experiences*, pages 110–111, New York, New York, USA, 2004.

[124] Yossi Rubner, Carlo Tomasi, and Leonidas J. Guibas. The Earth Mover's Distance as a Metric for Image Retrieval. *Int. J. Comput. Vision*, 40(2):99–121, 2000.

[125] Phillipe Salembier and Thomas Sikora. *Introduction to MPEG-7: Multimedia Content Description Interface*. John Wiley & Sons, Inc., New York, NY, USA, 2002.

[126] Daniel L. Schacter. *The Seven Sins of Memory: How the Mind Forgets and Remembers*. Houghton Mifflin, 2001.

[127] Daniel L. Schacter, Lana Israel, and Carrie Racine. Suppressing False Recognition in Younger and Older Adults: The Distinctiveness Heuristic. *Journal of Memory and Language*, 40(1):1–24, 1999.

[128] Daniel L. Schacter and Endel Tulving. *Memory Systems 1994*, chapter What are the memory systems of 1994, pages 1–38. MIT Press, Cambridge, MA, USA, 1994.

[129] Patrick Schmitz. Inducing Ontology from Flickr Tags. In *Collaborative Web Tagging Workshop at WWW2006*, Edinburgh, Scotland, 2006.

[130] R. Schweickert and B. Boruff. Short-term memory capacity: Magic number or magic spell? *Experimental Psychology: Learning, Memory, and Cognition*, 12:419–425, 1986.

[131] Abigail J. Sellen, Andrew Fogg, Mike Aitken, Steve Hodges, Carsten Rother, and Ken Wood. Do life-logging technologies support memory for the past?: an experimental study using sensecam. In *CHI '07: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 81–90, San Jose, California, USA, 2007.

[132] Hayri Sever and Mehmet R. Tolun. Comparison of Normalization Techniques for Metasearch. Izmir, Turkey, 2002.

[133] Mehmet Sezgin and Bulent Sankur. Survey over image thresholding techniques and quantitative performance evaluation. *Journal of Electonic Imaging*, 13(1):146–168, 2004.

[134] R.N. Shepard. Recognition memory for words, sentences, and pictures. *Journal of Verbal Learning and Verbal Behavior*, 6:156–163, 1967.

[135] Norman J. Slamecka and Peter Graf. The generation effect: Delineation of a phenomenon. *Journal of Experimental Psychology: Human Learning and Memory*, 4:592–604, 2007.

[136] Alan F. Smeaton. *ARIST - Annual Review of Information Science and Technology, Vol. 38, Chapter 8*, chapter Chapter 8. Indexing, Browsing and Searching of Digital Video, pages 371–407. American Society for Information Science and Technology, 2004.

[137] Alan F. Smeaton and Paul Browne. A Usage Study of Retrieval Modalities for Video Shot Retrieval. *Information Processing and Management*, 42(5):1330–1344, 2006.

[138] Alan F. Smeaton, Dermot Diamond D, Philip Kelly, Kieran Moran, King-Tong Lau, Deirdre Morris, Niall Moyna, Noel E. O'Connor, and Ke Zhang. Aggregating Multiple Body Sensors for Analysis in Sports. In *pHealth 2008 - International Workshop on Wearable, Micro and Nano Technologies for the Personalised Health*, Valencia, Spain, 2008.

[139] Alan F. Smeaton, W Kraaij, , and P Over. TRECVID 2003 - An Overview. In *TRECVID 2003 - Text Retrieval Conference TRECVID Workshop*, Gaithersburg, Maryland, 2003.

[140] Alan F. Smeaton and Paul Over. The TREC-2002 Video Track Report. In *TREC 2002 - Text Retrieval Conference*, Gaithersburg, Maryland, 2002.

[141] Alan F. Smeaton, Paul Over, and Wessel Kraaij. TRECVID: Evaluating the Effectiveness of Information Retrieval Tasks on Digital Video. In *12th ACM International Conference on Multimedia 2004*, pages 652–655, New York, NY, 2004.

[142] Alan F. Smeaton, Paul Over, and Ramazan Taban. The TREC-2001 Video Track Report. In *TREC 2001 - Text Retrieval Conference*, Gaithersburg, Maryland, 2001.

[143] Arnold W.M. Smeulders, Marcel Worring, Simone Santini, Amarnath Gupta, and Ramesh Jain. Content-based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22:1349–1380, Dec 2000.

[144] D. Tancharoen, T. Yamasaki, and K. Aizawa. Practical Life Log Video Indexing Based on Content and Context. In *Multimedia Content Analysis, Management, and Retieval In Proceedings of SPIE-IST Electronic Imaging*, San Jose, California, USA, 2006.

[145] Datchakorn Tancharoen and Kiyoharu Aizawa. Novel Concept for Video Retrieval in Life Log Application. In *PCM Pacific Rim Conference on Multimedia*, pages 915–923, Tokyo, Japan, 2004.

[146] Datchakorn Tancharoen, Toshihiko Yamasaki, and Kiyoharu Aizawa. Practical experience recording and indexing of Life Log video. In *CARPE '05: Proceedings of the 2nd ACM workshop on Continuous archival and retrieval of personal experiences*, pages 61–66, Hilton, Singapore, 2005.

[147] S. Tano, T. Takayama, M. Iwata, and T. Hashiyama. Multimedia Informal Communication by Wearable Computer based on Real-World Context and Graffiti. In *ICME IEEE International Conference on Multimedia and Expo*, pages 649–652, Toronto, Ontario, Canada, 2006.

[148] Endel Tulving. *Levels of Processing in Human Memory*, chapter Relation between encoding specificity and levels of processing, pages 405–428. Lawrence Erlbaum Associates, 1979.

[149] Endel Tulving. *Basic mechanisms in cognition and language*, chapter Neurocognitive processes of human memory, pages 261–281. Elsevier, 1998.

[150] Roelof van Zwol. Flickr: Who is Looking? In *WI: Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence*, pages 184–190, Silicon Valley, CA, USA, 2007.

[151] Faraneh Vargha-Khadem, David G. Gadian, and Mortimer Mishkin. Dissociations in cognitive memory: the syndrome of developmental amnesia. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 356(1413):1435–1440, 2001.

[152] Hal R. Varian. Economics and search. *SIGIR Forum*, 33(1):1–5, 1999.

[153] Sunil Vemuri and Walter Bender. Next-Generation Personal Memory Aids. *BT Technology Journal*, 22(4):125–138, 2004.

[154] Ellen M. Voorhees. TREC: Continuing information retrieval's tradition of experimentation. *Commun. ACM*, 50(11):51–54, 2007.

[155] Zhe Wang, Matthew D. Hoffman, Perry R. Cook, and Kai Li. VFerret: content-based similarity search tool for continuous archived video. In *CARPE '06: Proceedings of the 3rd ACM workshop on Continuous archival and retrival of personal experences*, pages 19–26, Santa Barbara, California, USA, 2006.

[156] Ziyou Xiong, Xiang Sean Zhou, Qi Tian, Yong Rui, and Thomas S. Huang. Semantic Retrieval of Video. *IEEE Signal Processing Magazine*, 23:18–27, Mar 2006.

[157] M.M. Yeung and Boon-Lock Yeo. Time-constrained clustering for segmentation of video into story units. *Pattern Recognition, 1996., Proceedings of the 13th International Conference on*, 3:375–380, Aug 1996.

[158] Jeffrey M. Zacks, Nicole K. Speer, Jean M. Vettel, and Larry L. Jacoby. Event Understanding and Memory in Healthy Aging and Dementia of the Alzheimer Type. *Psychology and Aging*, 21(3):466–482, 2006.