Investigating Keyframe Selection Methods in the Novel Domain of Passively Captured Visual Lifelogs

Aiden R. Doherty, Daragh Byrne, Alan F. Smeaton, Gareth J.F. Jones and Mark Hughes
Centre for Digital Video Processing and Adaptive Information Cluster,
Dublin City University, Glasnevin, Dublin 9, Ireland.
adoherty@computing.dcu.ie

ABSTRACT

The SenseCam is a passive capture wearable camera, worn around the neck, and when worn continuously it takes an average of 1,900 images per day. It can be used to create a personal lifelog or visual recording of the wearer's life which can be helpful as an aid to human memory. For such a large amount of visual information to be useful, it needs to be structured into "events", which can be achieved through automatic segmentation. An important component of this structuring process is the selection of keyframes to represent individual events. This work investigates a variety of techniques for the selection of a single representative keyframe image from each event, in order to provide the user with an instant visual summary of that event. In our experiments we use a large test set of 2,232 lifelog events collected by 5 users over a time period of one month each. We propose a novel keyframe selection technique which seeks to select the image with the highest "quality" as the keyframe. The inclusion of "quality" approaches in keyframe selection is demonstrated to be useful owing to the high variability in image visual quality within passively captured image collections.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval; H.5.0 [**Information Interfaces and Presentation**]: General

General Terms

Algorithms, Human Factors, Experimentation, Measurement

Keywords

Keyframe Selection, Visual LifeLogs, Image Quality Metrics

1. INTRODUCTION

The aim of lifelogging is to automatically capture and retrieve personal data encountered on a daily basis stored over a lifetime, e.g. web pages browsed, e-mail received, conversations, and images of activities participated in. Visual lifelogs, which include images from daily life, have received much attention recently [7, 3, 11]. The focus has now shifted from hardware miniaturisation and storage to that of managing and retrieving the stored information [1, 11]. In this paper we specifically work with the Microsoft SenseCam device [7], a wearable digital camera for passive capture of life experiences. This continuously captures a series of photos and similar to video content, the resulting visual lifelog can be segmented into discrete units or "events" [3]. Visual lifelogs of this type are by their nature extremely voluminous, typically growing by an average of 1,900 images per day per person (which equates approximately to 22 unique events per day). With such large, continuously growing collections, there is a significant information management challenge. As such the selection of appropriate keyframes to represent events becomes increasingly important. With the larger amounts of data, enabling a content owner to quickly and efficiently interrogate their generated content (or search results) is vital. Consequently, the keyframes selected to embody each event must be highly representative of that content and must convey its core concepts.

Of course the use of keyframes is not unique to the domain of lifelogging. Keyframes are ubiquitously used in video retrieval as a means by which an at-a-glance summary can be offered to users. Digital video content is typically 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 [2] note that *"keyframes must both represent the underlying video clip and distinguish that clip from the remainder of the collection"*. As such, an ideal keyframe accurately summarises the major concepts contained within a media segment allowing a user to quickly identify segments relevant to their information need.

The selection of keyframes for visual lifelog content is however not without challenges unique to the domain. First, the keyframe selection methods which have shown success and prevalence in the domain of video may not necessarily translate directly to visual lifelogs. For example, motion analysis [19] is an extremely popular mechanism for keyframe selection in video. Such a mechanism relies on the high frequency of video capture (i.e. 24/30 frames per second); however within passive capture lifelogs the rate of capture is variable (dependant on onboard sensors) and for the Microsoft Sense-

CIVR'08, July 7–9, 2008, Niagara Falls, Ontario, Canada.

Cam can be as low as 1 frame every 50 seconds. As a result, the visual information can vary widely from frame to frame making such motion analysis extremely difficult. Motion of the camera itself is however automatically captured by the Microsoft SenseCam using an onboard accelerometer, and this may offer an alternative source of motion information which we investigate for keyframe selection in the lifelogging domain.

In addition, passive capture devices may not always capture high quality images [6]. Unlike video where the frames which compose a shot are normally of a consistent high standard, within a lifelog the frames composing an event can vary widely in quality. For example, a quick short movement of the wearer at the time of image capture may result in extreme blurring of the frame. Other features such as obscuring of the image due to clothing, fingers or hands, and covering some or all of the lens, are also quite common. In an earlier study we found that a significant proportion (39%) of the images in a SenseCam lifelog collection are of poor quality owing to blurring, light saturation, overly dark conditions, or noise [6].

Finally, as a lifelog collection can be expected to grow daily, any keyframe selection and analysis technique must be efficient enough to deal with the large amount of content that is generated. Images are typically downloaded from the device on a daily basis and data should be available shortly afterwards to the owner.

In summary, visual lifelogs are extremely voluminous collections consequently their owners need to be able to rapidly interrogate them. The most appropriate mechanism to enable this is through a representative keyframe, however as outlined above, lifelogs are novel collections and methods from similar domains such as video do not necessarily translate in a straightforward manner. They additionally pose a significant challenge to choosing an appropriate keyframe due to the volume of images contained within a single event and the varied quality of those images. Thus, an investigation into keyframe selection methods is required.

In this paper we outline the findings of our investigation into keyframe selection within the domain of lifelogs. We present the findings from a detailed comparative evaluation of several potential selection methods for visual lifelogs and more specifically for the Microsoft SenseCam. We outline several possible methods by which keyframes can be selected from such a visual lifelog and experimentally investigate their success at selecting representative keyframes for lifelog events, given the challenges previously outlined. The results from our experiment are presented along with some recommendations for possible future work.

2. BACKGROUND AND RELATED WORK

2.1 Lifelogging and the Microsoft SenseCam

Recording of personal life experiences through digital technology is a phenomenon we are increasingly familiar with: music players, such as iTunes, remember the music we listen to frequently; our web activity is recorded in web browsers' "History"; and we capture important moments in our lifetime through photos and video [1]. This concept of digitally capturing our memories is known as lifelogging.

To enable increased non-intrusive capture of visual lifelog material, Microsoft Research in Cambridge, UK, have developed a device known as the SenseCam. The SenseCam

Figure 1: The Microsoft SenseCam

is a small wearable device that passively captures a person's day-to-day activities as a series of photographs [7]. It is typically worn around the neck, and so is oriented towards the majority of activities which the user is engaged in. Anything within the view of the wearer can be captured by the SenseCam. At a minimum the SenseCam will take a new image approximately every 50 seconds, but sudden changes in the environment of the wearer as detected by onboard sensors, can trigger more frequent photo capture. The device requires no manual intervention by the user as its on-board sensors detect changes in light levels, motion and ambient temperature and then determine when is appropriate to take a photo. For example, when the wearer moves from indoors to outdoors a distinct change in light levels will be registered and photo capture will be triggered.

The SenseCam takes an average of 1,900 images in a typical day, and as a result a wearer can very quickly build large and rich photo collections. Within just one week, over 13,000 images may be captured and over a year the lifelog photoset could grow to 675,000+ images. The benefits of this are numerous and include the ability for a user to easily record events without having to sacrifice their participation, aiding memory and recall and providing insight into a person's life and activities [1]. Notably, preliminary work between Microsoft Research and Addenbrooke's hospital in Cambridge, U.K indicates that a rich photo lifelog can dramatically improve memory and recall for individuals with neurodegenerative memory problems [7].

Recent investigations [6] into the composition of visual lifelogs indicate that they not only differ from traditional photosets in volume, the type of content captured, and the concepts contained in each image, but also in the quality of the images. The visual quality is extremely varied and lifelogs consequently tend to contain a large portion of unusable or sub-standard quality images (up to 40% of a collection). By implication, the authors recommend the filtering or removal of such images as the content tends not to be obvious from casual examination of such images. This is of particular relevance to keyframe selection within this domain.

2.2 Event Segmentation

Previous work [3, 11] recognised the need to automatically divide lifelog photosets into discreet events. This challenge is quite similar to that of scene boundary detection in video as opposed to shot boundary detection, as events or activities have an inherent underlying semantic meaning.

However given the nature of lifelog images (particularly the low-quality SenseCam images), it is unrealistic to expect the performance of segmenting events in the lifelogging domain to approach that of scene boundary detection for video data. Harder still will be to achieve the typically very high level of performance for shot boundary detection in video. In previous work we carried out an extensive evaluation to optimise event segmentation for lifelog images from the SenseCam [3]. The results recommended a system combining image features (content) and the accelerometer sensor values (context) giving an overall F-Measure value of 0.6237.

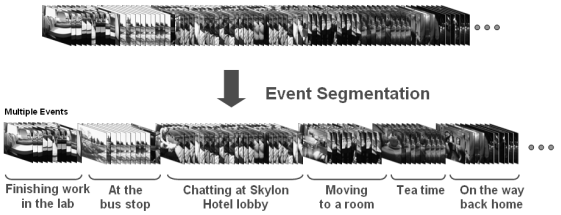

Figure 2: Overview of event segmentation

2.3 Keyframe Selection

Given that an event consists of many images, the challenge is to then select an appropriate representative keyframe image for each one. We are not aware of any related work within the domain of lifelogging, and little work has been reported even in the domain of video or image processing in terms of comparative evaluation of keyframe selection approaches. While there are several sophisticated approaches to keyframe selection within video retrieval, Smeaton & Browne [17] note that the simple approach of taking the middle frame is often favoured as the shot keyframe image [15, 5] due to its relatively good performance and computational efficiency. As such, we use this approach as the baseline against which to compare our other approaches. Cooper & Foote [2] investigated two other keyframe selection methods, namely: selecting the individual image that is closest to all the other images in the event; and 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 approaches can be computationally expensive though Kehoe & Smeaton [8] have explored taking advantage of graphic processor units (GPUs) to quickly and efficiently select keyframes using the first method of Cooper & Foote [2].

3. KEYFRAME APPROACHES

Before keyframes can be selected, event boundaries have to be determined and because of the volume of data this has to be done automatically. We now give details of our approach to detecting events, subsequent sections then describe our methods for determining keyframes; these use both traditional techniques and our proposed novel techniques incorporating the concept of image quality.

3.1 Event Segmentation

Figure 3 provides an overview of our system for segmenting a day's worth of images into distinct events or activities. Essentially our event segmentation approach attempts to identify periods of visual or sensory change, and identifies those occasions as most likely to be boundaries between distinct events or activities.

Firstly sequences of SenseCam images are broken up into a series of chunks, where the boundary between these chunks corresponds 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 images. Each image is then represented by MPEG-7 descriptor values and values from SenseCam sensors described earlier. The MPEG-7 descriptors used are: colour layout, colour structure, scalable colour, and edge histogram.

To segment a day of images into distinct events, processing follows these steps [3]:

- Compare adjacent images (or blocks of images) against each other to determine how dissimilar they are. A histogram intersection vector distance method is used to compare adjacent (blocks of) images.
- Determine a threshold value whereby higher dissimilarity values indicate areas that are likely to be event boundaries (mean thresholding [3] with $k = 3.4$)
- Post-processing: Remove successive event boundaries that occur too close to each other.

In previous work on a dataset of 271,163 images from 5 distinct users, who had manually groundtruthed 2,986 event boundaries, this approach recalled 62.17% of all boundaries (recall). 62.57% of all boundaries proposed by the system were valid boundaries (precision), resulting in an overall F-Measure score of 0.6237 [3].

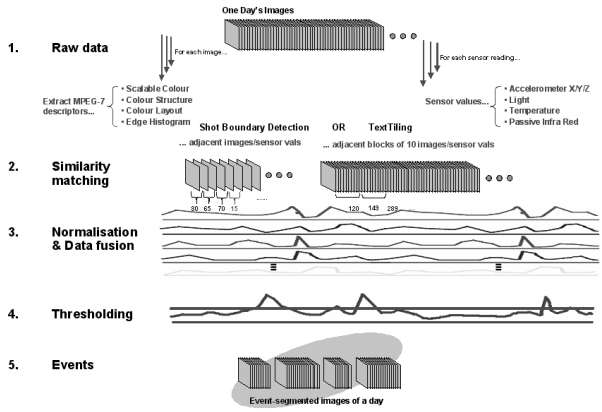

Figure 3: Overview of processing step to segment images into events

3.2 Traditional Keyframe Selection Techniques

Following segmentation we investigate three approaches to keyframe selection braodly similar to those used in other literature. The first approach is to simply select the middle image. Thereafter we investigate selecting the image that is most representative of a given event, and also the image that is most representative of a given event but also most different to the other events. There are a number of research challenges inherent to these approaches, as we describe in the following subsections. Using a small training set of 101

events (8,247 images) we now explain our answers to these important questions. All scores reported in the following subsections indicate the average score for a given approach across all 101 keyframe judgments on a 1-5 Likert scale.

3.2.1 Compare Images Exhaustively or Against Event Average?

To select an event representative image, Cooper & Foote select the image that is closest to all other images in the event (requiring $n * n$ comparisons, where n is the number of images in a given event) [2]. We investigate selecting the image that is closest to the average of all the other images in a given event, resulting in just n comparisons. A negligible difference in performance of user judgement on both approaches was found (3.35 vs. 3.33), with processing load significantly reduced using our approach.

Cooper & Foote also discuss another method to select the image that is closest to all the other images in a given event, but most different to all the other images in the other events [2]. We investigate the selection of the image that is most representative of an event by being closest to the average value of that event, but also what distinguishes it best from other events (by comparing against the average value of each of the other events). This reduces processing from $n * (n + m)$ 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 $m = 90 * e]$. Little difference was found between both approaches (3.01 vs 3.14) with processing load being vastly reduced (4% of time of other $n * (n + m)$ approach) in our proposed approach.

3.2.2 Best Vector Distance Metric

In the previous section it is required that we compare images against each other (or the event average vector), and therefore we investigated the Histogram Intersection (Likert score of 3.35), Kullback-Leiber (3.30), Manhattan (3.54) and Euclidean (3.59) approaches to image comparison. On our small training set the Euclidean approach performed best.

3.2.3 Increase Emphasis on Images Towards Middle of Event

Smeaton notes that *"... The danger of choosing from the start or end of a shot is the increased likelihood of picking up artifacts from the previous shot if there has been a gradual rather than a hard shot transition..."* [16]. Therefore we investigated if linearly weighting images towards the middle as being better candidates for a keyframe proved beneficial. This proved better in our training set (3.59 vs. 3.25), so we decided to opt for this approach.

3.2.4 Normalisation and Fusion of Data Sources

Sum and Max-Min normalisation were investigated for fusing data sources with Sum normalisation performing *marginally* better (3.59 vs. 3.45).

The data sources we used are unweighted as in our training set there was no clear advantage offered by assigning various confidences to the sources of information. We compared CombMED (score of 3.57), CombSUM (3.57), CombMIN (2.43) and CombMAX (3.53). Given that CombSUM is regularly chosen as the standard fusion method, we decide to use this approach too.

3.2.5 Weighting of Within vs. Cross Event

For the second approach discussed by Cooper and Foote [2] (section 3.2.1) there is a trade-off on the emphasis to place on how representative an image should be of the event it belongs to, as against how much it should be different from all the other images of the day. After investigating a number of different weights, it was decided to attach an equal weighting to both elements.

3.3 Image Quality Measures

Given that a visual lifelog can contain images of highly varied quality, it is likely that quality may play an important role in the selection of an appropriate keyframe. Quality measures were thus extracted automatically from each image within the collection. The five low-level image features described below were explored as a measure of image quality. The extraction of all features for a single image, takes approximately 1 second using a 2.3Ghz Intel Core 2 Duo machine with 2GB RAM. After processing, the values are aligned with the event, and normalised.

- **Contrast Measure**. Image contrast is a measure of the ratio of the intensity of the brightest color (white) to that of the darkest color within an image. A very low or very high contrast value indicates poor image quality, a median contrast measure is preferred. This is calculated by converting each pixel in the image from the RGB colour space to the YUV colour space. The image was then split into 8x8 image blocks. In each of these blocks the maximum Y value, which correlates to pixel intensity, and the minimum Y value were calculated. The minimum value was then subtracted from the maximum value to give a contrast value for the image block. The average of all these contrast values was then calculated to give an overall contrast measure for each image.

- **Colour Variance**. This is intended to correlate with the perception of colour richness. Since only the colour variance among the dominant colours in an image was desired, the colour space was divided into eight bins: black, white, red, green, blue, yellow, cyan and magenta. Each pixel value was examined and stored in appropriate bin using the smallest Euclidean distance between the respective colour values. The number of pixels in each bin was examined and compared against a threshold (empirically determined to be 20%). The variance of the colour values contained in the bins that are above this threshold was then calculated.

- **Global Sharpness**. This is intended to correlate with the perception of how sharply focused the image is. For this we wanted to measure the sharpness based only on sections of an image that were in focus. To calculate the sharpness measure we used a technique outlined in [13]. Edge detection is first performed on the image using the Sobel operator. In this case, each image block above a certain threshold is marked as an edge block. The average edge width is then calculated across all these edge blocks to give the overall sharpness measure.

- **Noise Measure**. Image Noise is a random, usually unwanted, fluctuation of pixel values in an image. The more noise within an image the lower the perceived

quality. To calculate the amount of noise in an image, we examined each pixel. The mean value was calculated for each pixel's 3x3 neighborhood and the Euclidean distance between each pixel value in the neighborhood and the neighborhood mean was calculated. If the original pixel at the center of the neighborhood has the maximum distance from the neighborhood mean, then this pixel is marked as noise. To calculate the overall measure of noise for each image, the amount of noisy pixels are added up and divided by the number of pixels in the image.

- **Saliency Measure**. This measure is intended to correlate with 'busyness' within an image 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. To calculate the saliency measure we select salient regions using the method described by Lowe [12], where scale-space peaks are detected in a multi-scale difference-of-Gaussian pyramid. The average of the values at these scale space peaks was then calculated and normalized to give us our overall saliency measure.

Additionally two measures were extracted from the SenseCam's data file for use. As these values are automatically captured by the SenseCam and potentially suppliment or replace the quality metrics above they were included in the examination of potential approaches to image quality judgment.

- **Accelerometer.** The Microsoft SenseCam contains an on-board XYZ accelerometer which provides readings of a wearer's motion. This reading was seen to be complimentary to the blur quality feature since it could be expected that if there are high levels of motion at the time of image capture, it is highly likely that the image will be blurred due to that motion.
- **Light Sensor.** The SenseCam additionally contains a light sensor, which measures the amount of ambient light around the wearer. As many low quality images result from light saturation or lack of light (darkness), this sensor could potentially be useful in identifying poor-quality images. In the case of this sensor, values should not tend towards either end-point (too dark or too bright)

3.3.1 Selecting A Quality Approach

With a starting point of possible measures for selecting a keyframe using image quality, several approaches were evaluated on a training collection prior to a full evaluation. The training collection consisted of 8,247 images from 101 events belonging to one user. In total 11 quality approaches to keyframe selection were explored. Within each approach a combination of the normalised quality measures (low-level image features and/or sensor readings) were fused using the CombSum technique. The highest scoring frame from each event for each approach was then selected and rated by a single annotator on a five-point Likert scale. The eleven approaches inspected were:

1. **Sensor Values:** An unweighted fusion of the sensor values (Accelerometer and Light) for each frame of an event;
2. **Basic Quality:** Unweighted fusion of Blur, Noise & Colour Variance;
3. **Weighted Approach 1:** A weighted fusion of Blur (0.2), Noise (0.2) and Colour Variance (0.6);
4. **All Quality Measures:** A fusion of all extracted quality measures for each frame (Blur, Noise, Colour Variance, Contrast, and Salience);
5. **All Quality & Sensor:** An unweighted fusion of all metrics for each image (Accelerometer, Light, Blur, Noise, Colour Variance, Contrast and Salience);
6. **Combination Approach 1:** An unweighted fusion of all quality measures except Blur (the accelerometer sensor values are used instead). This was used to determine the usefulness of the accelerometer values for judging blur when compared with approach 4;
7. **Combination Approach 2:** An unweighted fusion of all quality measures except contrast (the light sensor values are used instead). This was used to determine the usefulness of the light values for judging constrast when compared with approach 4;
8. **Simple Approach 1:** Unweighted fusion of Contrast and Salience;
9. **Simple Approach 2:** Unweighted fusion of Blur, Contrast & Salience;
10. **Simple Approach 3:** Unweighted fusion of Blur, Colour Variance, Contrast & Salience
11. **Weighted Approach 2:** A complex weighted fusion of Blur and Noise (0.25) further combined with a fusion of Constrast and Salience (0.75)

Table 1 shows the performance of the quality approaches for keyframe selection within the training collection in terms of an assessor's rating on a 5-point Likert scale (higher is better). Five approaches yield very promising results (approaches 4, 7, 8, 10 & 11) however approaches 7 and 8 perform best. The accelerometer shows little promise in indicating blur or poor quality images since its inclusion over Blur shows a substantial drop in performance (approach 6 vs. approach 4). However, the use of light in determining quality shows some potential with the mild increase of approach 7 over 4.

In further comparison of the results we contrast the performance of various approaches against the average performance of unique keyframes within the same event. Again approaches 7 and 8 outperform other approaches. Both performed better than average 67.33% of the time. However, approach 7 only performed equal to the average in 14.85% of events with approach 8 equaling average performance 16.83% of the time. Given that approach 8 mildly outperforms approach 7, and also given its relative simplicity and computational efficiency, it was favoured for evaluation over other methods of keyframe selection in visual lifelogs.

3.4 Approaches for Investigation

Based on the initial analysis carried out in this section on various keyframe selection techniques and approaches to determining image quality, we selected the following methods for an extensive evaluation of their usefulness as keyframe selectors:

	Approach	Average Score
1	Sensor Values	2.91
2	Basic Quality	2.21
3	Weighted Approach 1	2.72
4	All Quality Measures	3.67
5	All Quality & Sensor	3.40
6	Combination Approach 1	3.42
7	Combination Approach 2	3.72
8	Simple Approach 1	3.72
9	Simple Approach 2	3.49
10	Simple Approach 3	3.67
11	Weighted Approach 2	3.70

Table 1: Average performance of various quality approaches against test collection

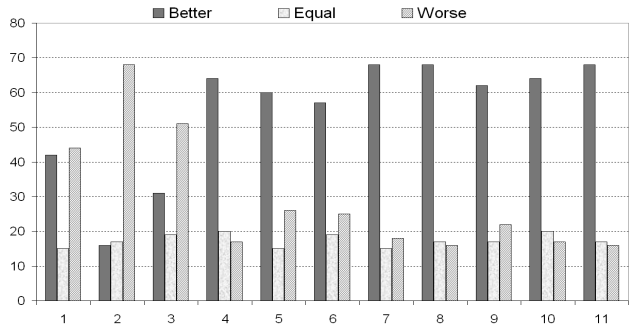

Figure 4: Comparision of performance as either better than, equal to or worse than the average performance of unique keyframes selected

- **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 the 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, that also has a good image quality, and is finally distinguishable from the images in the other events;

4. EXPERIMENT OVERVIEW

After deriving 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 do this, keyframes were selected for lifelog events using the various approaches and subsequently judged manually. The details of the experimental evaluation are presented in Table 2.

User	Days	Images	Events	Judgements Made	No Judge Required
1	35	25,243	360	1,323	837
2	44	67,755	790	2,761	1,985
3	21	42,693	408	1,409	1,045
4	25	40,681	491	1,681	1,271
5	9	18,485	183	639	459
Total	134	194,857	2,232	7,813	5,597

Table 2: Distribution of 2,232 events over 5 users

A number of people within our research group have worn the Microsoft SenseCam device. However, only a small number have worn the device over an extended period, continuously recording their life experiences. Four such individuals (in their early twenties to mid thirties) participated in our keyframe selection experiments. In each case 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. 99 days worth of visual lifelog data were used in the experiments, equating to almost two hundred thousand images (see Table 2).

For each image within the experimental collection, the contrast and salience quality measures were extracted automatically as described earlier. The collection was then segmented into 2,232 discrete events using our segmentation method [3]. Potential keyframes for each event were then selected using the various methods. 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 six potential keyframes per event.

The owners of the original SenseCam 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 for lifelog data they had originally generated. As the same frame will offer the same overview of the concepts contained within an event we were able to significantly reduce the number of judgments required by each participant. In the case of the same frame being selected by more than one method, the participants judged that frame only once. This resulted in a reduction of 5,597 judgments (see Table 2) and ensured consistency in the judgments i.e. the same frame could not be rated differently as the single judgment applies to all selection approaches.

In order to facilitate the keyframe judgment process, a custom tool was developed (see Figure 5). The tool was installed on each participant's desktop computer. Each user completed the judgments at their leisure. The application provided feedback to the user as to their current progress through the task. At launch or when a judgment was completed, the application selected at random a keyframe to annotate from the pool of remaining un-judged frames. Additionally, a loading message was displayed for 2.5 seconds between judgments. The random selection and presentation delay were introduced to mitigate against priming and interaction effects.

While making a judgment, the keyframe under scrutiny

was 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' recollection of the event. Every eighth image in the event was presented to provide a summary of the event and as a comparative set by which to judge the proposed keyframe.

Users were provided with a set of radio buttons below the keyframe and event images. Users rated the keyframe by directly 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.

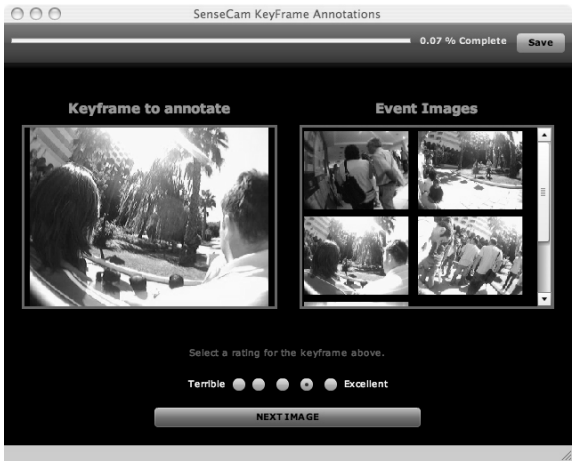

Figure 5: Annotation tool used for experiments

5. RESULTS AND DISCUSSION

Judgements for 13,410 keyframes were provided by 5 users. Ratings provided for each approach were analysed and are presented in Table 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.

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 a fusion of both offers 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 that they are known to be composed of many poor quality images with high variance in quality over short periods of capture [6]. This variation present within events explains the effectivness of the quality measures in 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.

Approach	Avg. Likert Score (higher = better)
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 3: Overall performance on entire dataSet

Figure 6 illustrates the performance of each approach across the users' collections. It can be observed that there is a mild variation in the levels to which users have assessed the performance of each approach, however they are largely consistent with the overall findings. In all cases a combination of image quality and either "*Within Event*" or "*Cross Event*" outperforms the baseline. Interestingly, we see that quality was not as effective in the collection of user 3, but was very effective for users 1 and 5. This would indicate that the quality feature performance is variable when used independently, dependent on the collection. However, this collection effect appears to be tempered by fusing it with other measures.

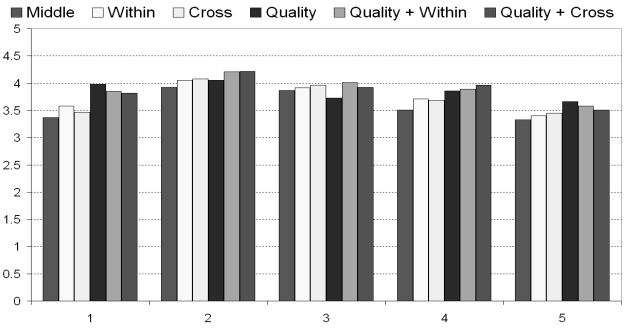

Figure 6: Performance of approaches for each user

Lifelog images are typically downloaded and processed on a daily basis. Additionally, a review of lifelog data typically involves providing a browseable daily summary to the user. 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. Figure 7 demonstrates that again combination of image quality with either "*Within Event*" or "*Cross Event*", significantly outperforms other approaches for most day's events. Both approaches prove to be at least as good as the baseline 80% of the time (in terms of overall performance within all the events in a day) with both offering better performance over 69% of the time. Additionally most approaches generally perform better than the baseline. It is noteworthy that the quality measure alone as a keyframe selection approach, while performing well overall, does not appear to provide as consistently good keyframes at the daily level.

More detailed exploration of the "Within Event and Image Quality Fusion" approach's performance for each day's worth of events is provided in Figure 8, which contrasts this

approach with the baseline for each day across all users.

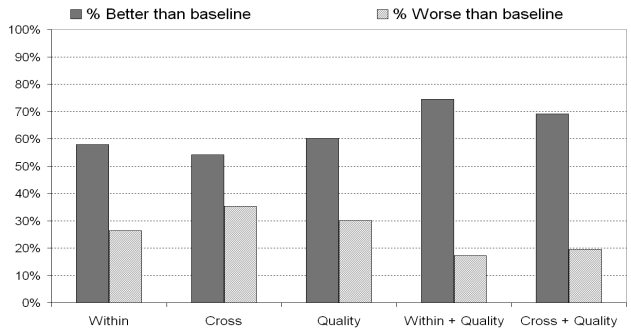

Figure 7: Overall daily performance improvement offered by each approach over middle image

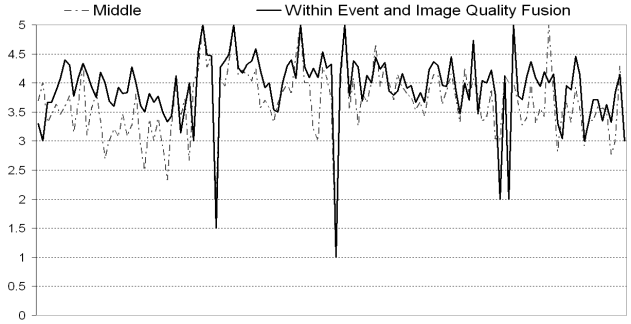

Figure 8: Performance of Quality+Within vs Baseline Across All Days

5.1 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 challenges include the limitations of event segmentation, events with a high proportion of visual change, and events with high visual change as a result of the nature of the activity. Given the difficulties in selecting keyframes, we decided to explore the performance of the various approaches in events containing a high amount of visual change.

5.1.1 Issues relating to event segmentation

The best performing systems in the last TRECVid scene boundary detection task achieve an F1-Measure approaching 70%[9], which is in all likelihood the best that could be expected with visual lifelog segmentation. In fact even this level of performance should not be expected since the SenseCam has a fisheye lens which makes the comparison of adjacent (low-quality) images particularly challenging.

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 select a representative keyframe. Participants in the judgment effort indicated as much in anecdotal feedback following their annotation.

5.1.2 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 its frames. As images are captured every 22 seconds (on average) major changes in the visual landscape can appear from frame-to-frame within an event. This is particularly 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 particularly challenging to select an image that neatly represents and summarises the salient concepts of the event. Conversely, with 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.

Figure 9: Example of a highly variable visual event

5.2 Selection of 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 standard deviation of the first bin of the colour layout feature was calculated for all the images in each event in the test collection. After training on a set of 369 events from two users, a standard deviation threshold value of 13 was selected. Within this training set, it should be noted that the first user only achieved a precision score of 0.48 (66/113) for correctly 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.

5.3 Performance of Approaches on Events With High Visual Variability

The six keyframe selection approaches were then investigated for only those events above the visual change threshold (Table 4). This comprised a 337 events out of 2,232 total events available. We can see that the scores for events with a high degree of visual variability are notably lower than the

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 4: Overall performance of each approach on events with high image variability

reported performance for *all* events (see Table 3). This confirms that there is indeed a significant challenge in selecting keyframes for such events. Quality measures perform well here as these events are likely 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 10. 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, again in the case of user 3 all approaches work almost equally well, while in the case of users 1 and 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.

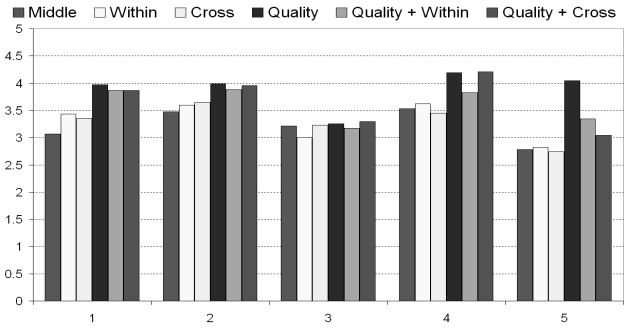

Figure 10: Performance for each user on events with high image variability

Comparison of the approaches (against the baseline) within each day's worth of (high image variability) events (shown in Figure 11) highlights the consistent performance of both the quality and quality combination approaches. Across all days and only taking events with high inter-image variability into consideration, the approach that improves most on the baseline is the "*Image Quality*" approach. This approach works 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 11 (only events with high inter-image variability) to those of Figure 7 (*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 it does across *all* events. All 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 12.

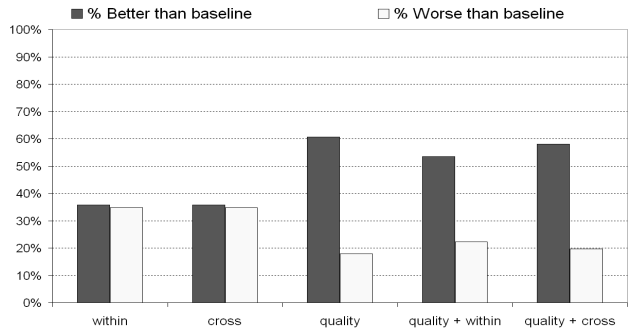

Figure 11: Improvement offered by each approach over a number of days of high variability events

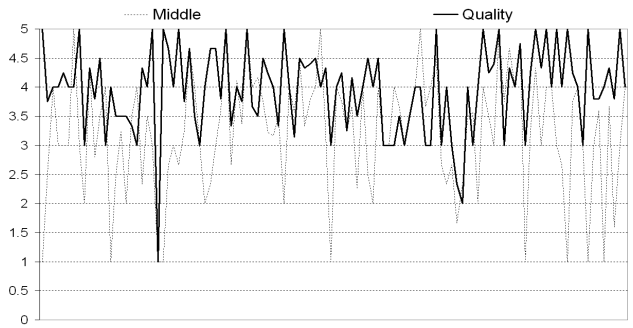

Figure 12: Performance for each day on events with high image variability

6. CONCLUSIONS

The challenge of selecting a relevant keyframe from a lifelog event is particularly acute and has motivated the comparative evaluation of approaches presented in this paper. Individual events in the lifelogging domain can vary greatly in terms of visual quality, with up to 40% of images being blurred, noisy, light-saturated, very dark, etc. This presents a particular challenge that traditional keyframe identification techniques struggle to adequately address. In this paper we proposed a technique to select keyframes based on the image within an event that has the best image "quality". One drawback of our proposed approach is the extra computational overhead, however across 69.92% of all days the "*Image Quality*" approach returns a set of keyframes (considering all the events in each day) at least as good as the standard approach of selecting the middle image, and performs 6.07% better on average overall. However we identified that there are a large number of events in the lifelogging domain that can vary greatly in terms of image quality, and just taking those events into consideration the "*Image Quality*" approach performs at least as good as the baseline on 82.14% of days, and 15.48% better on average overall.

7. FUTURE WORK

Our investigation has explored several methods for keyframe selection suited to visual lifelogs, but it is not complete. As lifelog collections often contain a wide range of both content and context data there is the potential for other novel approaches to keyframe selection. As part of future investigation into keyframe selection for lifelog collections we suggest exploring some of the following approaches.

Within a lifelog **biometric recordings** can be bound to other sources of data (including the events within visual lifelog collections) to provide emotional and affective context for the data [14]. Times of increased emotional intensity may indicate good instances at which to select an emotionally significant or "affective" keyframe.

Bluetooth, the short range communications technology prevalent on mobile devices is ideally suited to supplementing a SenseCam visual lifelog with **co-presence information**. Changes in the number, proximity and the social importance or "familiarity" [10] of co-present devices to the owner could provide cues to significant points in an event. We expect that this would offer a means to further extend the approaches examined in this paper.

Concept detection is used frequently in video retrieval to extract semantic concepts from frames of digital video footage [18]. By matching the visual features of a frame within the footage to the properties of known "concepts" (such as indoors, outdoors, people, crowd, etc.) the probability of a concept's occurrence within the frame is determined. With concept detection enabled, keyframes could be selected based on the number, prevalence and distribution of (interesting) concepts within frames of the event. Again, we expect that this would offer a means to further extend the approaches examined in this paper.

Finally, in previous work we have investigated using **face detection** with respect to SenseCam images [4]. Although recall performance was low (0.2872), the precision was encouraging (0.6336). We intend to evaluate the selection of keyframe images based on the presence of faces.

Acknowledgements: We would like to extend our thanks to the participants in these experiments. We are grateful to the AceMedia project for equipment. This work is supported by Microsoft Rearch under grant 2007-056; the Irish Research Council for Science Engineering and Technology; and by Science Foundation Ireland under grant 03/IN.3/I361.

8. REFERENCES

[1] G. Bell and J. Gemmell. A digital life. *Scientific American*, 2007.

[2] M. Cooper and J. Foote. Discriminative techniques for keyframe selection. In *ICME 2005 - IEEE International Conference on Multimedia and Expo*, 2005.

[3] A. R. Doherty and A. F. Smeaton. Automatically segmenting lifelog data into events. In *WIAMIS 2008 - 9th International Workshop on Image Analysis for Multimedia Interactive Services*, 2008.

[4] A. R. Doherty and A. F. Smeaton. Combining face detection and novelty to identify important events in a visual lifelog. In *CIT 2008, Workshop on Image- and Video-based Pattern Analysis and Applications*, 2008.

[5] E. Dumont and B. Merialdo. Split-screen dynamically accelerated video summaries. In *TVS 2007 - TRECVID BBC Rushes Summarization Workshop, ACM Multimedia 2007*, 2007.

[6] C. Gurrin, A. F. Smeaton, D. Byrne, N. O'Hare, G. J. Jones, and N. O'Connor. An examination of a large visual lifelog. In *AIRS 2008 - Asia Information Retrieval Symposium*, 2008.

[7] S. Hodges, L. Williams, E. Berry, S. Izadi, J. Srinivasan, A. Butler, G. Smyth, N. Kapur, and K. Wood. SenseCam: A retrospective memory aid. In *UbiComp Ű 8th International Conference on Ubiquitous Computing*, 2006.

[8] P. Kehoe and A. F. Smeaton. Using graphics processor units (GPUs) for automatic video structuring. WIAMIS - 8th International Workshop on Image Analysis for Multimedia Interactive Services, 2007.

[9] W. Kraaij, A. F. Smeaton, and P. Over. Trecvid 2004 - an overview. In *TRECVID 2004 - Text Retrieval Conference TRECVID Workshop*, MD, USA, 2004. National Institute of Standards and Technology.

[10] B. Lavelle, D. Byrne, C. Gurrin, A. F. Smeaton, and G. J. Jones. Bluetooth familiarity: Methods of calculation,applications and limitations. In *MIRW - Workshop at MobileHCI07: 9th International Conference on HCI with Mobile Devices*, 2007.

[11] W.-H. Lin and A. Hauptmann. Structuring continuous video recordings of everyday life using time-constrained clustering. In *Multimedia Content Analysis, Management, and RetievalŰ SPIE-IST Electronic Imaging*, 2006.

[12] D. Lowe. Distinctive image features from scale-invariant keypoints. In *International Journal of Computer Vision*, volume 20, pages 91–110, 2003.

[13] P. Marziliano, F. Dufaux, S. Winkler, and T. Ebrahimi. A no-reference perceptual blur metric. *Image Processing. 2002. Proceedings. 2002 International Conference on*, 3, 2002.

[14] C. Mooney, M. Scully, G. J. Jones, and A. F. Smeaton. *Investigating Biometric Response for Information Retrieval Applications*, pages 570–574. Springer, Berlin / Heidelberg, Germany, 2006.

[15] M. J. Pickering, D. Heesch, R. O. Callaghan, S. Ruger, and D. Bull. Video retrieval using global features in keyframes. In *TREC 2002 - Text Retrieval Conference*, MD, USA, 2002. National Institute of Standards and Technology.

[16] A. 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.

[17] A. F. Smeaton and P. Browne. A usage study of retrieval modalities for video shot retrieval. *Information Processing and Management*, 42(5):1330–1344, 2006.

[18] C. G. Snoek, M. de Rijke, G. Schreiber, and M. Worring. Adding semantics to detectors for video retrieval. *Multimedia, IEEE Transactions on*, 9(5):975–986, Aug. 2007.

[19] W. Wolf. Key frame selection by motion analysis. In *ICASSP '96: Proceedings of the Acoustics, Speech, and Signal Processing*, pages 1228–1231, Washington, DC, USA, 1996. IEEE Computer Society.