Approaches for Event Segmentation of Visual Lifelog Data

Rashmi Gupta and Cathal Gurrin*

Insight Centre for Data Analytics, Dublin City University, Dublin, Ireland
rashmi.gupta3@mail.dcu.ie, cgurrin@computing.dcu.ie*

Abstract. A personal visual lifelog can be considered to be a human memory augmentation tool and in recent years we have noticed an increased interest in the topic of lifelogging both in academic research and from industry practitioners. In this preliminary work, we explore the concept of event segmentation of visual lifelog data. Lifelog data, by its nature is continual and streams of multimodal data can easy run into thousands of wearable camera images per day, along with a significant number of other sensor sources. In this paper, we present two new approaches to event segmentation and compare them against pre-existing approaches in a user experiment with ten users. We show that our approaches based on visual concepts occurrence and image categorization perform better than the pre-existing approaches. We finalize the paper with a suggestion for next steps for the research community.

Keywords: $\cdot Lifelogging \cdot EventSegmentation \cdot FeatureExtraction \cdot MemoryAugmentation \cdot InformationRetrievalSystem.$

1 Introduction

Lifelogging is concerned with capturing and utilization of rich volumes of personal behavioural/activity data from multimodal lifelogs, gathered by individuals, who may be termed lifeloggers. These lifeloggers could be researchers or any individual who wish to capture the totality of their life [15]. Lifelog data could be collection of images, audios, videos, text documents and/or biometric data gathered using various wearable devices (e.g. wearable cameras) or software sensors. Lifelogging provides detailed information about the activities of the individual and could help to change an individual's behaviour so as to achieve positive life benefits. A variety of lifelog devices have been available with the Microsoft SenseCam [4], as used in MyLifeBits project [2], being the most well known. In addition, many other wearable sensors exist such as wearable cameras, biometric sensors, physical activity sensors, etc. can together passively contribute to a rich media digital diary which captures a representation of the individual's life activities. One aspect of such lifelog archives is that they tend to be passively captured and continuous (streamed) in nature [13], hence there exists a challenge in segmenting these continuous content streams into index-able units for

analysis, retrieval and presentation. Most retrieval systems are based on the core concept of a document as an indexing unit. In lifelog search and retrieval, the document is not clearly defined, due to the continuous nature of lifelog data, and efforts have been made to impose a unit of retrieval, such as the minute [14] or the event [8], which is a document-centric unit.

In this paper, we propose two new approaches for event segmentation of visual lifelog data using a dataset of 14,132 images from 10 users over the period of 1 day each (12-14 hours). These new approaches to segmentation are based on visual analysis of the visual image stream to identify objects and activities as a source for segmentation. An example of the types of data streams and their associated activities are shown in Figure 1. The contributions of this paper are: (i) the introduction of two new approaches for event segmentation of visual lifelog data, and (ii) a dataset and evaluation approach for evaluating event segmentation approaches for visual lifelog data.

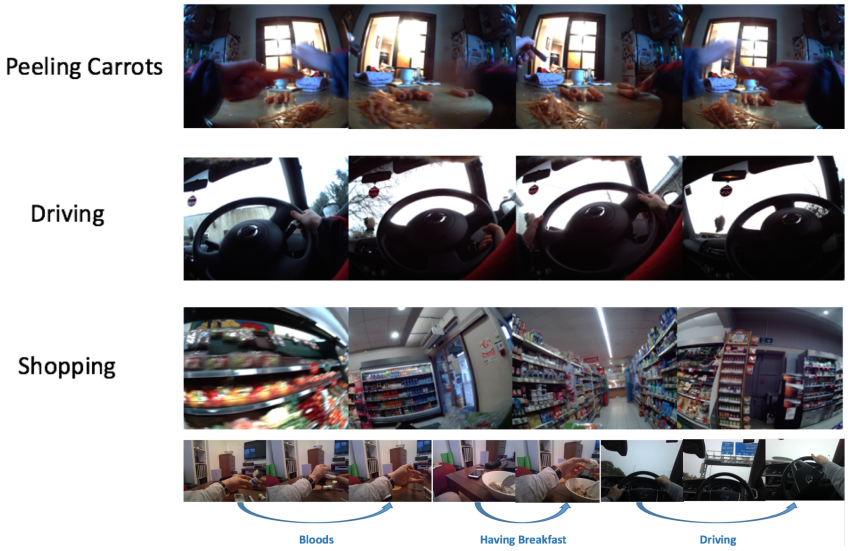

Fig. 1. Example of segmented daily life activities in lifelog dataset and signifying the transition between different activities.

2 Background

Lifelog data is typically based on passive capture of an individual's life experience. The data generated by lifelogging tends to be multimodal in nature, and streamed (as opposed to bursty in nature). Lifelogging has a long history, tracing back to Richard Buckminster Fullers Dymaxion Chronofile [7], in which he physically recorded all his personal and business data in a chronological arrangement

as a very large scrapbook. Steve Mann in 1980s, introduced the idea of digitally capturing continuous everyday life data with wearable computing and streaming videos. Later in 2006, Bell and Gemmell introduced "MyLifeBits", a software database of Bell's life [12]. Following this initial work in digitally recording daily life, there has been an increase in the availability of wearable sensors such as cameras (OMG Autographer, Narrative clip, iOn SnapCam, etc), fitness trackers (Fitbit, FuelBand, Jawbone wristband), smartphones apps (Moves, Saga), various biometric sensors, and more recently informational sensors, such as loggerman [16] which capture all computer interactions of an individual. We note that the process of capturing such rich volumes of digital multimedia data by the individual is becoming a normative activity. Recent years has seen the proliferation of visual capture devices such as cameraphones and digital video recorders, such as GoPros. What makes the content created by such devices differ from lifelog content is that they tend to produce bursty content. A cameraphone for example takes conventional photos in sequences of one (or more) at various times throughout the day. These datasets are naturally segmented into events or points in time based on the gaps between data capture. Whereas in lifelogging, the data streams are continuous and there is no clear point of segmentation. Consider an individual wearing a modern wearable camera. Such devices are usually worn attached to clothing or on a lanyard around the neck and can 'observe' the activities that the individual configured to capture images. Hence, we need to consider how to segment these data streams. In lifelogging, this segmentation process creates a contiguous set of documents that have typically been combined into a logical unit called an 'event'in a process called 'event segmentation'.

2.1 Event Segmentation of Lifelog Data

Event segmentation refers to the process whereby a continuous stream of data (typically from sensors) is segmented into discrete units. Zacks and Tversky in 2001, define the event as "*a segment of time at a given location that is conceived by an observer to have a beginning and an end*" [19]. Initial work on image-based event segmentation resulted from the ready availability of personal photo data from cameraphones. One early approach to segmentation of the bursty photo capture stream in timestamped data is discussed by Gargi in 2003 [11], which models the data stream with poisson distribution and used box-counting method.

In lifelogging, this automatic segmentation into events is similar to segmentation of video into shots and scenes and requires structuring the personal data into discrete units [15] which can be semantically enriched to form the basis of a lifelog retrieval system. Event segmentation of lifelog data has received research interest for about a decade now, yet there has not been much effort put into comparative evaluations. Doherty et al. [9] in 2007, implemented event detection for Sensecam image data by representing each image by a low-level edge histogram, a scalable colour (global), Color Histogram in HSV Color Space, accelerometer values of the Sensecam device and temperature readings as a source of evidence for the segmentation process. To determine the similarity between adjacent blocks of images, Hearst's TextTiling Algorithm was used. Following his early work,

Doherty [8] in 2008 introduced an enhanced event segmentation algorithm for wearable camera data using visual MPEG-7 features from images, which lead to an improvement over the previous approach. Various vector distance methods implemented and this work showed that the histogram intersection method and euclidean distance method based on MPEG-7 features perform best. In addition, kapur and mean thresholding approach were used as optimal thresholding techniques. Byrne et al. [5], presented an event segmentation technique based on content (using five low-level MPEG-7 feature descriptors) and contextual information (with light sensor i.e changes in light and human motion sensor i.e change of location/motion) of the lifelog image set using bluetooth and GPS metadata.

Chen et al. in 2011, gathered a large dataset of 450,000 images, about 2,000 hours of computer activities and 18 months of context data (350,000 records) from 3 lifeloggers [6]. The fusion of this rich lifelog data is segmented based on computer activities, location and visual concepts using a TextTilling algorithm. Li et al. in 2013 [18], employed event segmentation based on multi-sensor data recorded by a wearable camera with associated gyroscope and accelerometer data. To generate the event boundaries the S-STD (sum of all standard deviations) feature is extracted from gyroscope data and to fine-tuned to enhance performance. Additionally Segment-HSV (the mean of HSV histograms) feature is utilized. Most recently, a segmentation approach based on unsupervised hierarchical agglomerate clustering was introduced by Bolanos et al. [3] and evaluated over a small dataset of 4,005 images (part of three people's days).

To conclude the previously implemented experiments to segment continuous visual data, the researchers used different volumes of image lifelog data and in some cases fused visual data with other sensors. They extracted various types of visual features from images and implemented clustering techniques such as hierarchical agglomerative clustering to segment daily life activities of the day in to specific events. In this work, we implement new approaches to event segmentation based on high-level visual concepts and categories and evaluate these against a baseline approach represented by Doherty [8]. In addition, we also define an evaluation methodology that provides a repeatable and fair comparison between the different approaches.

3 New Approaches to Event Segmentation of Lifelogs

Although there are a number of approaches to event segmentation that we could take, we have chosen to compare high-level modern visual features with the baseline low-level computer vision based features used by Doherty [8]. We employ two high-level sources, the open-source CAFFE framework (1000 ImageNet classes) concept detector [17] and the image categorization detector (86-categories Taxonomy) provided by the Microsoft Computer Vision API (MS) [1], which is based on [10]. The process of segmenting one day visual lifelog data into meaningful events is shown in Figure 2 (below). The wearable camera that we used generated about two images per minute and these are organised into basic minute-long

atomic units (1440 minutes/day) by selecting first image of each minute; visual concepts are extracted using one of two approaches (outlined below); the continuous data stream is segmented into events and then the output is evaluated.

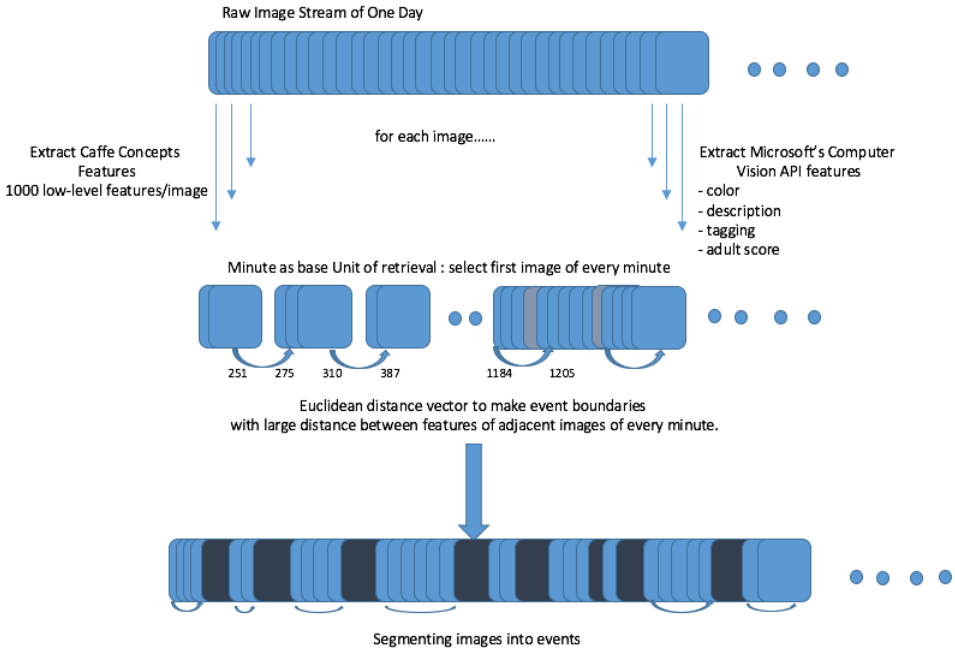

Fig. 2. Process of segmenting one day lifelog visual achieve into events.

3.1 Event Segmentation based on Visual Concepts

In the visual concept approach to event segmentation, it is our conjecture that the change of activities of an individual would result in a change in visual objects in the field of view of the individual. Therefore, we employed the Caffe framework [17] to detect the objects visible in lifelog image content. Caffe is deep learning framework, used in conjunction with 1,000 ImageNet dataset of visual concepts [17]. Hence, the Caffe visual concepts form a 1,000 item vector for each image. The process of event segmentation of one day lifelog images based on caffe concepts in specific events by observing activity change is shown in Figure 3.

3.2 Segmentation based on Image Categorization

The aim of this approach is to utilize a higher-level semantic categorization of the images as a source of evidence for the segmentation process. Microsoft's

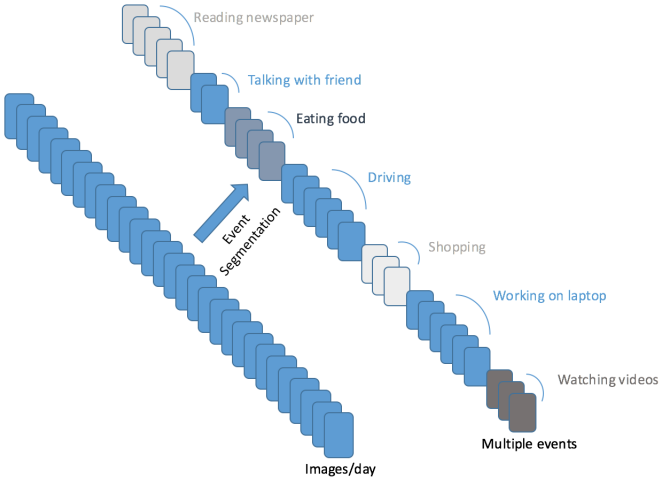

Fig. 3. Event segmentation of one day images based on Visual Concepts.

Computer Vision API [1] is employed for this task; it returns a taxonomy-based categorization for each image into 86 semantic categories. These categories are organised as a taxonomy into parent/child hierarchies. This taxonomy includes different categories, such as indoor category (includes indoor_churchwindow, indoor_door, indoor_room,etc.), food category (includes food_grilled, food_bread, food_pizza,etc.), outdoor category (includes outdoor_mountain, outdoor_city) and so on. This results in one vector (of size 86) representing each lifelog image with the associated confidence values. The example of identified various categories in image content and segmented images of one day based on these category taxonomy is shown in Figure 4.

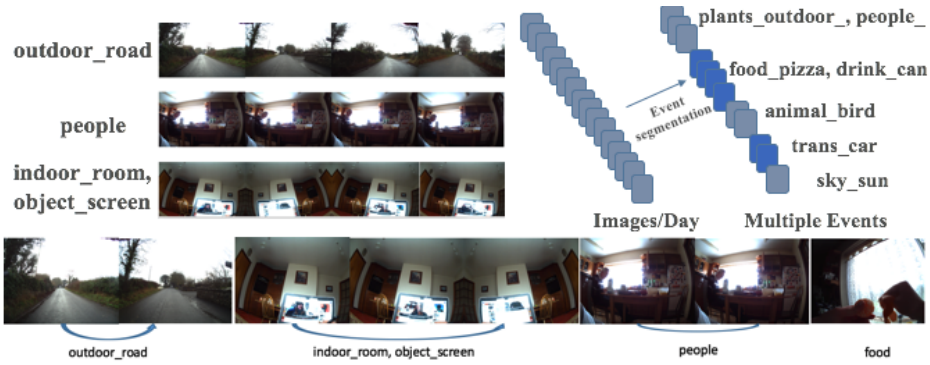

Fig. 4. Example of segmented lifelog dataset based on categories.

3.3 Baseline Approach

In order to compare against pre-existing approaches, we developed a baseline approach based on the work of Doherty [8]. This segmentation approach is based on, MPEG-7 low-level visual feature extraction from SenseCam images, TextTiling (block of 5 adjacent images) and Non-TextTiling approaches, various distance measures and threshold determination techniques. The approach to segment visual lifelog data implemented by Byrne [5] is similar to the Doherty [8] approach and performs similarly. As a consequence, for our baseline approach, we re-implemented only the Doherty approach.

3.4 Distance Measure

A key component of event boundary detection is the ability to identify the distance between subsequent lifelog images (or groups of images) and where the distance is above a certain threshold (along with other criteria), an event boundary can be declared. In order to calculate the distance, we implemented the euclidean distance measure on these vectors, which allows us to identify the event boundaries (a high distance) and fuse the images within the same event (low distance). Therefore a change in activities with high distance such as 'eating' and 'driving' would be highly distant from each other, so they would indicate an event boundary, whereas activities such as 'eating' and 'cooking' would have a lower distance and may not trigger an event boundary. This requires the selection of appropriate thresholds, and this is described in the following section.

3.5 Threshold Determination

To identify the most effective events boundaries, we need to determine the threshold values. We implemented, two automatic thresholding techniques based on pre-existing approaches; the first is a parametric technique (e.g. mean thresholding which takes mean, standard deviation and user parameters) and second is non-parametric technique (Kapur thresholding). The manual thresholding parameters we selected were 0.4, 0.5, 0.6 and 0.7, which can be subject to more fine tuning at a later date.

3.6 Avoiding Over-Segmentation

As with prior work, we needed to avoid over-segmenting the data, which could happen if there is significant visual change in a short sequence of camera images, which can commonly occur in lifelogging due to short-term variations in the activities of the individual. We propose that such small variations are not representative of changes in the overall activities of the individual, and as such, we should not segment based on these. Hence we chose five minutes as the smallest duration of a segmented event, which is chosen based on our experience of analyzing and organizing lifelog data.

Lifelogger/User	Profession	Age group	Avg duration/-day	Avg images/day	Total ground truth events/day
1	Researcher	>40	17 hours	1084	46
2	Researcher	>35	12 h 35 m	1792	33
3	Researcher	>35	13 hours	1895	16
4	Researcher	>35	17 hours	1078	31
5	Researcher	>25	11 hours	1064	31
6	Student	>30	11 hours	1399	20
7	Student	>22	11 hours	1494	39
8	Student	>25	14 hours	1336	35
9	Businessman	>55	13 hours	1708	41
10	House-maker	>45	13 hours	1282	19

Table 1. The summary information of participants and their count of segmented ground truth events.

4 Evaluation

The automatic event boundaries generated by the two proposed approaches and the baseline approach are compared with manual event segmentation done by the lifeloggers themselves which is considered as ground truth event segmentation. For this evaluation we developed a new purpose-built visual lifelog dataset gathered by a number of individuals.

4.1 Dataset

In this experiment, we collected total 14,132 images (average of 1,400 images over the period of 11 to 14 hours per day/each user) from 10 different participants. The OMG Autographer was used for data capture, which is a passive-capture wearable camera worn on a lanyard around the neck and therefore is oriented towards the activities of the wearer. Typically the camera will capture about 2 images per minute. All other sensors on the device, such as bluetooth and GPS were turned off to optimize battery life. All the participants are asked not to change their daily routine due to wearing the camera. Each participant manually segmented their day into a set of discrete activities, which we then take as a ground truth event segmentation. We provided the same guidance to each participant on the process to be employed when deciding what should be considered to be an event boundary. As a consequence, there is a natural variability in the number of events manually segmented due to human subjective judgments being made and the numbers of events are in line with what we would have expected. The detailed information regarding participants and the average count of images collected and segmented from each user per day is summarized in Table 1.

4.2 Evaluation Methodology

One challenge when evaluating event segmentation algorithms is how to evaluate systems that produce a segmentation that is accurate, but labels the segmentation point to be a few images before, or after, the user-defined ground truth

segmentation point (close-to, but not exactly matching the subjective human segmentation). To solve this problem, we reuse the approach of Doherty (i.e. post-processing boundary gap) which defines a sliding window around the ground truth labeled segmentation-point. The size of this window could range from 0 (no sliding window allowed) up to an arbitrary figure of 16 (the upper-bound for reasonable experimentation [1]). Through experimentation, we have found that five images is a reasonable size of the sliding window around the ground truth labeled event boundary points. Every nearest boundary is considered to be true in the associated window and we used this boundary in our experimentation, explained in Figure 5 below.

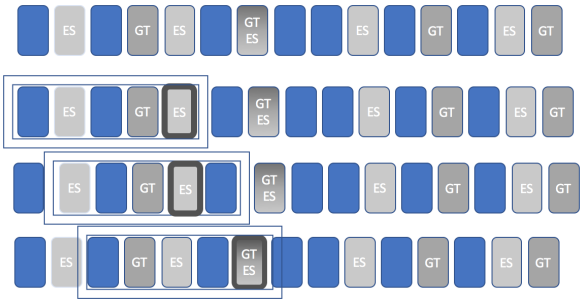

Fig. 5. Similarity between Event Segmentation Points (ES) and User-defined Ground Truth Event Segmentation (GT) via Sliding Window Approach (Post-processing Boundary Gap) [8].

In terms of evaluation measurements, we used the conventional measurements of precision, recall, F1-Measure and Matthews Correlation Coefficient (MCC). Precision and recall are the standard approaches to evaluation measurement in information retrieval. F1 score is harmonic mean of precision and recall and MCC score does not only taken into consideration for correct prediction but also measures the correlation between all values in a matrix and identifies the mis-predictions by adopting values smaller than 0.

4.3 Results

The two new approaches introduced in this paper are compared with the baseline approach discussed in section 3 in a comparative study.

– **Baseline: Event Segmentation based on MPEG-7 Descriptors:** The two pre-existing segmentation approaches implemented by Doherty [8] is

[1] Given a 16 hour day, with one image per minute and 30 events identified per day (all reasonable assumptions), then an evaluation with a 16 minute boundary would tend not to penalize random segmentation algorithms.

based on, TextTiling (Block of 5 adjacent images) and non-textiling are implemented. We get the highest score of precision (20.6%), recall (65.8 %), F1-Measure (65.8%) and MCC (7.42 %) with hearst's textiling approach and with the non-textiling approach, we get the highest score of precision (29.5 %), recall (60.6 %), F1-Measure (38.7 %) and MCC (22 %) as shown in Table 2. We found the Non-textiling approach with mean thresholding technique performs best.

- **Event Segmentation based on Visual Concepts:** The approach to segment one day lifelog data into activities by using caffe framework visual concepts, as described in (section 3.1), performs better from MPEG - 7 low level features. With threshold value 0.4, we get best score of precision (70.4%), F1-Measure (69.3%) and MCC (64.3%) shown in Table 2 below.
- **Event Segmentation based on Image Categorization:** Segmentation of one day lifelog data into events based on the image categories, as described in (section 3.2), we get the best evaluated scores of recall (68.3 %), F1-Measure (70.1 %) and MCC score (65.7 %) with threshold value 0.7 shown in Table 2, which again justifies the better approach over baseline approach with MPEG-7 Descriptors.

Experimental Approaches	**Threshold Value**	**Precision**	**Recall**	**F1- Score**	**MCC**
Hearst's TexTiling based on MPEG-7 Descriptors	mean(K = 0.5)	20.6	65.4	30.7	6.98
	Kapur	**20.6**	**65.8**	**31.3**	**7.42**
Non TexTiling based on MPEG-7 Descriptors	**mean(K = 0.5)**	**29.5**	**60.6**	**38.7**	**22.0**
	Kapur	29.4	60	38.7	21.43
Caffe Visual Concepts	**0.4**	**70.4**	**72**	**69.3**	**64.3**
	0.5	64.6	76.5	68.5	62.9
	0.6	56.4	80.9	64.8	58.6
	0.7	40.5	88.2	54.2	46
Image categorization via MS Concepts	0.4	78.3	65.5	69.2	65.2
	0.5	77.5	66.2	69.4	65.4
	0.6	77.2	67.2	69.8	61.4
	0.7	**76.2**	**68.3**	**70.1**	**65.7**
Summary Results	**Threshold Value**	**Precision**	**Recall**	**F1- Score**	**MCC**
MS Concepts	0.7	76.2	68.3	70.1	65.7
Caffe Concepts	0.4	70.4	72	69.3	64.3
MPEG-7 without TexTiling	mean (K = 0.5)	29.5	60.6	38.7	22.0
MPEG-7 with TexTiling	Kapur	20.6	65.8	31.3	7.42

Table 2. Overall Thresholding Performance based on MPEG - 7 Descriptors [8], Caffe Visual Concepts [17], Image Categorization via MS Concepts [1] and Summary of Experiment Results.

4.4 Discussion

As can be seen from the previous section, we found that the segmentation based on image categorization and visual concepts provide higher-level semantic concepts that reflects the differences in the activities of the individual. For example,

moving from the office desk to eat lunch, the visual concepts and objects [1,17] in the field of view would naturally change from computers to food items. Assuming this to be the case, then it is natural that a higher-level visual analysis would perform better than one that operates just on lower-level visual features such as edge histogram and colours of the MPEG-7 library [8]. Hence, we found that the event segmentation based on image categorization [1] with precision (76.2 %), recall (65.5 %), f1-score (70.1 %) and MCC (65.7 %) and event segmentation based on visual concepts [17] with precision (70.4 %), recall (72 %), f1-score (69.3 %) and MCC (64.3 %) provides the best results while re-implementation of baseline approach [8] with Non-TexTiling approach (found best approach in baseline) provides comparatively low results summarized in Table 2 above.

We note that these are preliminary approaches and are subject to optimisations and enhancements. We intend to explore more optimal thresholds as well as larger datasets [14].. The simple distance measure that we employed can also be enhanced and we intend to explore a number of alternatives, such as the wordnet-based conceptual distance, as well as the results of fusing many different approaches. This work is also limited by the fact that we only analyse the visual content. We began this paper by stating that lifelogs are multimodal data archives, so we will employ multimodal data sources for our future work, such as audio, acceleration, location, biometrics, etc.

5 Conclusion

In this preliminary work, we presented two new approaches for event segmentation of visual lifelog data based on high-level visual feature analysis. In order to place our research in the context of past state-of-the-art, we defined a baseline approach to segment visual lifelog data into retrievable events based on the work of Doherty et. al. [8]. We compared our two proposed approaches to the baseline approach in an experimental setting with the lifelog data of ten users. In this experiment, we showed that the higher-level approaches proposed in this paper perform significantly better than [8] across all four employed evaluation measures. This suggests that there is significant scope for enhancing the performance of event-segmentation algorithms on lifelog data and that this is far from a solved problem. In future, we plan to extend this work along the lines previously outlined. We also intend to compare our approaches with the full spectrum of alternative approaches as introduced above.

Acknowledgment

This publication has emanated from research conducted with the financial support of Science Foundation Ireland (SFI) under grant number SFI/12/RC/2289.

References

1. Microsoft's computer vision api. `https://docs.microsoft.com/en-us/azure/cognitive-services/computer-vision/category-taxonomy`.

2. G. Bell and J. Gemmell. A digital life. *Scientific American*, 296:58–65, February 2007.
3. M. Bolanos, R. Mestre, E. Talavera, X. G. Nieto, and P. Radeva. Visual summary of egocentric photostreams by representative keyframes. *IEEE First International Workshop on Wearable and Ego-vision Systems for Augmented Experience (WEsAX)*, 29 June - 3 July, 2015, Turin, Italy.
4. V. Bush. As we may think. *The Atlantic Monthly*, 176(1):101–108, Jul 1945.
5. D. Byrne, B. Lavelle, A. R. Doherty, G. J. Jones, and A. F. Smeaton. Using bluetooth and gps metadata to measure event similarity in sensecam images. *5th International Conference on Intelligent Multimedia and Ambient Intelligence*, July 2007.
6. Y. Chen, G. J. Jones, and D. Ganguly. Segmenting and summarizing general events in a long-term lifelog. *In: The 2nd Workshop Information Access for Personal Media Archives (IAPMA) at ECIR 2011*, April 2011.
7. H. y. Chu and R. G. Trujillo. New views on r. buckminster fuller. pages 6–23, 2009.
8. A. R. Doherty and A. F. Smeaton. Automatically segmenting lifelog data into events. *9th International Workshop on Image Analysis for Multimedia Interactive Services*, 30 June 2008.
9. A. R. Doherty, A. F. Smeaton, K. Lee, and D. P. Ellis. Multimodal segmentation of lifelog data. *Large Scale Semantic Access to Content (Text, Image, Video, and Sound)*, pages 21–38, June 2007.
10. H. Fang, S. Gupta, F. Iandola, R. Srivastava, L. Deng, P. Dollar, J. Gao, X. He, M. Mitchell, J. Platt, L. Zitnick, and G. Zweig. From captions to visual concepts and back. IEEE Institute of Electrical and Electronics Engineers, June 2015.
11. U. Gargi. Modeling and clustering of photo capture streams. *Proceedings of the 5th ACM SIGMM international workshop on Multimedia information retrieval*, pages 47–54, November 2003.
12. J. Gemmell, G. Bell, and R. Lueder. Mylifebits: A personal database for everything. pages 88–95, January, 2006.
13. C. Gurrin, D. Byrne, N. E. O'Connor, G. J. Jones, and A. F. Smeaton. Architecture and challenges of maintaining a large-scale, context-aware human digital memory. *VIE 2008 - The 5th IET Visual Information Engineering 2008 Conference*, July 2018.
14. C. Gurrin, H. Joho, F. Hopfgartner, L. Zhou, and R. Albatal. NTCIR Lifelog: The First Test Collection for Lifelog Research, 2016.
15. C. Gurrin, A. F. Smeaton, and A. R. Doherty. Lifelogging: Personal big data. *Foundations and Trends in Information Retrieval*, 8(1):1–125, 2014.
16. Z. Hinbarji, R. Albatal, N. E. O'Connor, and C. Gurrin. Loggerman, a comprehensive logging and visualisation tool to capture computer usage. *22st International Conference on MultiMedia Modelling (MMM 2016)*, January 2016.
17. Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014.
18. Z. Li, Z. Wei, W. Jia, and M. Sun. Daily life event segmentation for lifestyle evaluation based on multi-sensor data recorded by a wearable device. *Conf Proc IEEE Eng Med Biol Soc*, 30, October 2013.
19. M. J. Zacks, S. T. Braver, A. M. Sheridan, I. D. Donaldson, Z. A. Snyder, M. J. Ollinger, L. R. Buckner, and E. M. Raichle. Human brain activity time-locked to perceptual event boundaries. *In: Nature neuroscience*, pages 651–655, 2001.