Visual Lifelogs Retrieval: State of the Art and Future Challenges

Zenonas Theodosiou[1*] and Andreas Lanitis[1,2†]

[1]Research Centre on Interactive Media, Smart systems and Emerging Technologies (RISE), Nicosia, Cyprus
[2]Dept. of Multimedia and Graphic Arts, Cyprus University of Technology, Limassol, Cyprus

Email: *z.theodosiou@gmail.com, †andreas.lanitis@cut.ac.cy

***Abstract*—The use of wearable cameras covers several areas of application nowadays, where the need for developing smart applications providing the sustainability and well-being of citizens it is more necessary than ever before. The tremendous amount of lifelogging data to extract valuable knowledge about the every day life of the wearers requires state of the art retrieval techniques to efficiently store, access, search and retrieve useful information. Several works have been proposed combining computer vision and machine learning techniques to analyze the content of the data captured from visual wearable devices on a daily basis. This paper presents an overview of the progress in visual lifelogging retrieval and indicates the current advances and future challenges, highlighting the prospects of incorporating visual lifelogging retrieval in social computing applications.**

***Index Terms*—wearable cameras, lifelogging, retrieval, digital memory**

I. Introduction

Within the general theme of ambient intelligence, wearable computing constitutes an important research and technological direction. Wearable devices are on the rise in recent years due to technological developments and play a significant role in Internet of Things (IoT) applications and in the highly promising future developments of ubiquitous computing. Wearable devices are electronic components integrated on clothing or accessories which can be easily worn from the users. Thus, technology companies have shown great interest by investing a lot in the development of innovative small components with embedded advanced sensing technologies to easily collect and transmit data from the wearer's environment.

Usually wearable cameras are small and light devices which can be fastened at human body covering the point of view of the wearer. They provide the capability to seamlessly record visual data in a passive way, in a first-person persepctive, while the wearer is performing her/his activities. A lot of research has been carried out using wearable cameras the last years, including studies related to: cognition and social interaction between humans [1], navigation/assistive technologies for the blind [2], monitoring and assistance of physical environments [3], automated life story creation [4], summarization [5], action (e.g. fall detection) or location recognition [6], security, safety and protection of citizens.

978-1-7281-3634-9/19/$31.00 ©2019 IEEE

Visual lifelogging is the seamless collection of images and/or videos through the use of wearable cameras and involves the continuous recording of the daily life of the wearer for a long periods of time [7]. The new field of the computer vision which deals with the content analysis of data collected by wearable devices, is called Egocentric Vision or First-person Vision. The analysis of such visual data can be successfully used to study everyday life and draw useful conclusions about human behavior, aiming to improve the quality of life and prevent people from mental disorders. Furthermore, visual lifelogging can be used as a digital memory to help elderly people suffering from memory disorders cope with the demands of modern lifestyle [8].

The enormous increase in the number of available visual lifelogging data requires the development of technologies for efficient archiving and access to visual content. Retrieving data from a digital collection can solve several problems in the field of Egocentric Vision, including: (1) searching for elements, (2) navigating, (3) understanding the environment, and (4) organizing huge amounts of data [7]. Due to the fact that the retrieval approaches which have already proposed for other type of data are not effective enough when applied for lifelogging tasks, there is a need for the development of beyond the state of the art techniques to successfully retrieve data from large scale lifelog databases.

In this paper we restrict our study on the current state of the art of visual lifelloging retrieval and demonstrate the potential of using lifelogging in social computing. In section II, the state of the art of the visual lifelogging retrieval is discussed. In section III, we present the techniques of analyzing the lifelog visual content. In section IV, various retrieval systems are investigated. The applications of the lifelogging retrieval are discussed in section V, followed by discussion and future challenges in Section VI. Finally, section VII concludes the paper.

II. State of the art

One of the main applications of visual lifelogging is to play the role of a visual memory where the lifelogs are saved in a database providing users with the opportunity to access the past memory anytime based on a query (Fig. 1). The successful management of huge amount of lifelogs needs

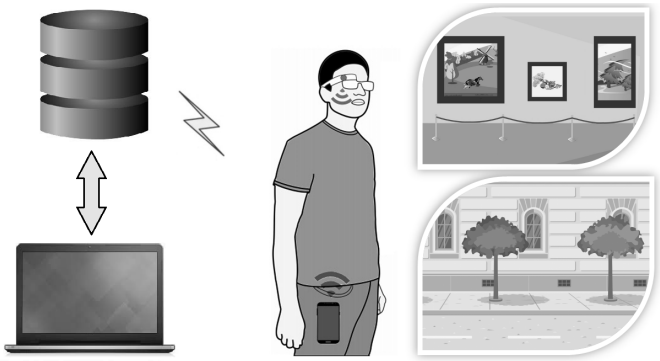

Fig. 1. Recording and access to past memories.

semantic indexing and retrieval. Retrieval can be performed either by content-based [9] or text-based methods [10]. The content-based approach performs the retrieval by examining the collection of images and returning those with similar visual content to the image given to the user's query. On the other hand, text-based approach returns images which are accompanied by text similar to the user's text query. Text-based image retrieval remains the predominant choice, despite the successful development of several content-based multimedia retrieval platforms. For effective retrieval using text-based techniques, annotation process is an essential step.

Annotation bridges the semantic gap and allows users to retrieve the visual content of their preference using text-based queries, relieving them of finding the image to be given as an example for content-based retrieval. Annotation can be manual or automatic by analyzing and understanding the visual content. The process of annotating lifelogs consists of detecting and identifying people and faces, activities occurring in events, locations, and date and time [11]. Lifelogging data usually contains repetitive information, so summarization appears a solution for improving the comprehension and avoiding the analysis of hundreds of images with same content. Thus, summarization is of great interest in the research community [7].

The analysis of the lifelog visual content requires reasoning about the shapes and materials in scenes, where the appearance of objects is modulated by illumination. Thus, the image classification (either refers to object recognition or activity detection) is very important step for image understanding. The process of classifying images using feature representation is a very demanding procedure due to the large number of visual differences between the images of the same class. Several classification studies have been conducted using handcrafed visual features aiming to overcome the above difficulties.

The first retrieval system was presented in 2007 by Gemmell et al. [25]. MyLifeBits software provides users the opportunity to easily search and retrieve lifelogging data. Since then, several approaches have been proposed including the collaborative benchmarking exercises (i.e. NTCIR, ImageCLEF) and challenges as the recent Lifelog Search Challenge.

III. Analysis of Visual Content

The traditional algorithms which utilize temporal coherence and motion estimation do not perform well when applied on data from wearable cameras in the wild. Recognition algorithms have to deal with the broad range of different objects presented in lifelogging data. In addition, the available data which is rapidly growing requires the development of efficient and effective techniques for analyzing the visual content according to the requirements of each application.

While low-level features (SIFT, SHIK, HOG, MPEG-7, etc.) [14] have been manually designed and successfully used in several classification and recognition tasks, they cannot have the same success when used in lifelogs creating the need of designing more sophisticated ones capable to cope with these type of data. The huge amount of data streams coming from visual lifelogging devices challenges the traditional machine learning approaches and calls the application of deep learning algorithms. According to the traditional machine learning flow, a feature extraction algorithm is applied to the input image and then the extracted features are used to create the classification model using a machine learning algorithm such as SVM, Decision Trees, Nave Bayes, etc.

The technological progress in computer industry has improved the time for training large neural networks and has revolutionized the field of computer vision [15]. Convolutional Neural Networks (CNN) are widely used in computer vision with great success including recognition and detection tasks [16]. The excellent performance of CNN in image analysis and understanding has led to the deployment of several products and services for computer vision applications in our daily lives.

The output of CNN layers can be interpreted as visual features and the algorithm plays simultaneously the roles of feature extractor and classifier. The excellent transfer learning properties of CNN allow the use of a feature representation learned for a specific task in a slightly different scenario. The top-layer or lower-layer [17] activations can be used as a feature representation of images for other tasks. Features extracted from pretrained CNN on ImageNet and ILSVRC13 datasets can be used as a feature extractor in computer vision tasks, including scene recognition and object detection, obtaining better performance results than the manually designed features [18].

IV. Retrieval Systems

The mechanisms of each lifelogging retrieval system differ according to the aim of use and type of archived data. Several systems have been proposed for retrieving indoor and outdoor everyday activities in large video archives the last decade using context-based and text-based approaches.

Wang et al. [20] presented the Vferret System for content-based searching and browsing of continuous archived video. The system gives users the opportunity to perform content-based similarity search using visual and audio features and combine it with other traditional search methods. The captured video is segmented into short video clips before inserted

into the system and then, video attributes, visual and audio features are extracted. During the retrieval step, an initial timeline-based filtering is performed and a k-means clustering algorithm clusters the results into a set of clusters so that a representative video clip is created for each cluster. The search results can be refined using content-based similarity between the results of the previous step and the clip query given by the user. The results of the evaluation showed that the combination of visual and audio features improves the search results compared to the results when using only the visual or audio features.

Aghazadeh et al. [21] tried to solve the problem of measuring the similarity of recorded background with the actual captured video. To this end, they proposed a system for the automatic detection of novel events in the life of the bearer who repeats a daily activity. A novel event is extracted when a query sequence is not registered to already saved sequences recorded while doing the same daily activity. Sequence registration is performed by calculating appearance and geometric similarity of individual frames and utilizing the invariant temporal order of the activity.

A retrieval method for large-scale egocentric visual data using a sparse graph representation was proposed in [23]. An object and scene retrieval from egocentric data is demonstrated using graph-based representation and clustering. The visual experiences are represented as a sparse graph where each node is an individual frame in the video and local density-based clustering algorithm is used for clustering the data showing that popular global clustering methods, like spectral clustering and Graclus, perform poorly on egocentric graph data.

A lifelog selection using semantic characteristics was presented in [24] which proposed a smart keyframe selection using summarization methods, people and semantically-rich scenes. The method is based on two parallel steps: (1) selection of keyframe using semantic segmentation and summarization of the whole day taking into account the common relationship of images, and (2) selection of image processes using the presence of faces and basic objects. The data retrieval is accomplished using text-based, inverted index retrieval methods combining the feature representations received from the advanced convolutional neural networks.

Ding et al. [26], proposed a method for generating video captions using long video segmentation, keyframe filtering and language model. Initially the redundant frames are detected and removed using the spatio-temporal interest points (STIPs) algorithm. The long video is segmented based on non-linear combination of different visual elements. Keyframes from the most impactful segments are converted to video captioning by using the saliency detection and LSTM variant network.

The selection of semantic concepts and their automatic detection was investigated in [22] to satisfy the requirements for searching, browsing or summarization of everyday multimedia lifelogs data. A semantic density-based algorithm which includes reasoning on semantic networks is proposed to select the most useful concepts to represent everyday lifelogging activities. The everyday activities which are related with lifelogging are selected using the criteria of time dominance, generality and high frequency. Various semantic similarity measures are investigated on two mainstream semantic networks, i.e. WordNet and ConceptNet, in order to represent the concepts and concept relations within lifelogging domain.

More recently, six interactive retrieval systems have been participated in the LSC2018 and tested on the same visual lifelog dataset where the four of them performed comparatively well, finding results within the time-limit for most of the topics. Some of the systems were based on existing retrieval systems [27], [28], [29] and other were developed to cover the needs of the challenge [30], [31], [32].

V. Applications

Visual lifelogging can be considered as a digital memory on behalf of the wearer holding her/his daily life (24/7/365) providing the opportunity of search, retrieve and share of its content. Visual lifelogs offer considerable potential for inferring knowledge about e.g. behavior patterns, and hence enable many applications that would not be possible with High Temporal Resolution (HTR) cameras related to self-monitoring, memory assistance, assisted living, as well as applications to infer conclusions for specific populations of users. Retrieval can provide the opportunity of accessing the digital memory of the wearer. Continuous access to digital memory could cover several needs which should be in mind while developing a new lifelog retrieval system. By incorporating the psychological aspects of memory in the design of such systems, five elements of human memory access should be considered to define a framework for lifelogging search and retrieval techniques [12]. The five aspects called the five R' s contains *Recalling* and *Recollecting* of the previous actions to help memory-augmentation applications, *Reminiscing* about previous actions to help personal wellness, *Reflecting* on previous actions for self-enhancement and *Remembering Intentions* for context-aware memory-assistance [13].

The lifelogging retrieval can play significant role in the improvement of autobiographical memory assisting people with poor short-term memory recall capacity. Therefore, several applications can be developed to improve the well-being of people with memory impairments [24], such as Alzheimer and other dementias.

The use of wearable cameras in the wild can significantly contribute to social progress and prosperity providing a broad range of indoor and outdoor applications. The idea of incorporating innovative systems into peoples everyday lives for continuous recording, storage and retrieval according to queries of specific purposes, could positively impact their lives by improving the living standards. Among others, visual lifelogging retrieval advances could be used in applications for people with vision or mobility impairments, city navigation, daily exercise and sports, education, museums and cultural heritage sites, citizens protection and security, elderly care, health rehabilitation, etc. The nature of wearable cameras which may record personal moments without the involved peoples' consent (e.g. use in public, etc.), requires the design

of applications that should be developed following the relevant legal and ethical frameworks. Therefore, the appropriate feasibility study should be conducted prior to the development of each application involving all the related public and private bodies to comply with General Data Protection Regulation (GDPR).

VI. Future Challenges

Initially the traditional information retrieval techniques used to manage lifelogging data provide access using text-based or content-based queries. However, the large amount of lifelog data requires more sophisticated retrieval mechanisms. The key challenges that visual lifelogging retrieval will face in the future are related to big volume of lifelog data which makes the state-of the art techniques unworkable and the main issues which multimedia indexing and retrieval region is facing.

The accurate visual content analysis is very important for successful retrieval. Although much progress has been made in computer vision, automated visual analysis of lifelogging data is still an open issue due to [19]: 1) bad image quality which suffers from blurring and artifacts owing to the not tight placement of the camera on the wearers body, (2) abrupt lighting variations, severe occlusions and non-informative images owing to the non-intentional nature of the data capturing (e.g. capturing of non-meaningful content like walls, details from objects, etc.), (3) minimum number of annotated datasets which causes several limitations in machine learning frameworks.

The above issues affect the performance of state of the art classification schemes, including deep learning algorithms, which give better results when they are trained with large annotated datasets. Furthermore, the ever-growing list of potential applications of image interpretation from wearable optical sensors, opens up new problem-specific challenges that require novel methodologies.

The creation of manually labeled datasets is time consuming and costly task and faces the limitations of the human subjectivity which is larger in the case of the lifelog data because of the low resolution and small size of the objects. On the other hand, the automatic lifelog labelling has to tackle the limitations of semantic gap and bridge the visual features with more vivid and diverse semantics in complex images.

In regard to retrieval mechanisms as other applications, there is a necessity of ranked retrieval results to queries of specific purposes. Extending this kind of retrieval mechanisms by incorporating access due recommendations using historical and immediate context sources, combined with multimodal and omnipresent interaction, will be one of the major advances in user access to lifelog data [11].

VII. Conclusions

Visual lifelogging can play the role of a digital memory on behalf of the wearer by recording the everyday activities and can play a significant role in the development of social applications. In this paper we present the current state of the art advances of visual lifelogging from the retrieval perspective which allows users to access their digital memory from anywhere and at anytime. We discussed the visual content analysis techniques used to detect useful information in lifelog data such as faces, concepts and activities using both low-level features and traditional machine learning algorithms as well as deep learning techniques. We summarized the systems which have been proposed so far to retrieve visual lifelog data and presented the possible applications of such systems. Finally we sum up the future challenges of the domain in regard to the big volume of lifelog data and retrieval mechanisms.

Acknowledgements

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement No 739578 (RISE) and the Government of the Republic of Cyprus through the Directorate General for European Programmes, Coordination and Development.

References

[1] M. Aghaei, M. Dimiccoli, and P. Radeva, "With whom do I interact? detecting social interactions in egocentric photostreams", CoRR, abs/1605.04129, 2016.

[2] T. Yagi, K. Mangalam, R. Yonetani, and Y. Sato, "Future Person Localization in First-Person Videos", arXiv preprint arXiv:1711.11217, 2017.

[3] S. R. Edmunds, A. Rozga, Y. Li, E. A. Karp, L. V. Ibanez, J. M. Rehg, and W. L. Stone, "Brief Report: Using a Point-of-View Camera to Measure Eye Gaze in Young Children with Autism Spectrum Disorder During Naturalistic Social Interactions: A Pilot Study", Journal of Autism and Developmental Disorders, vol. 47(3), pp. 898–904, 2017.

[4] Z. Lu and K. Grauman, "Story-driven summarization for egocentric video", In CVPR, 2013.

[5] N. Jojic, A. Perina, and V. Murino, "Structural epitome: a way to summarize ones visual experience", In Advances in neural information processing systems, pp. 1027–1035, 2010.

[6] A. Perina, M. Zanotto, B. Zhang, and V. Murino, "Location recognition on lifelog images via a discriminative combination of generative models", in Proc. of the British Machine Vision Conference. BMVA Press, 2014.

[7] M. Bolanos, M. Dimiccoli, and P. Radeva, "Toward storytelling from visual lifelogging: An overview", IEEE Trans. Human-Machine Systems, vol. 47(1), pp. 77–90, 2017.

[8] A.R. Doherty, S. E. Hodges, A. C. King, A. F. Smeaton, E. Berry, C. J. Moulin, and C. Foster, "Wearable cameras in health", American journal of preventive medicine, vol. 44(3), pp. 320–323, 2013.

[9] A. Nazir, R. Ashraf, T. Hamdani N. Ali, "Content based image retrieval system by using HSV color histogram, discrete wavelet transform and edge histogram descriptor", in Proc. International Conference on Computing, Mathematics and Engineering Technologies (iCoMET),2018.

[10] A. Dutta, Y. Verna and C. V. Jawahar, "Automatic image annotation: the quirks and what works", Multimedia Tools and Applications, vol. 77 (24), pp. 3199132011, 2018.

[11] C. Gurrin, A. F. Smeaton, and A. R. Doherty, "LifeLogging: Personal Big Data, Foundations and Trends in Information Retrieval, vol. 8(1), pp.1107, 2014.

[12] A.J. Sellen and S. Whittaker, "Beyond total capture: a constructive critique of lifelogging, Comm. ACM, vol. 53(5), pp.7077, 2010.

[13] R. Gupta, "Considering documents in lifelog information retrieval, In Proc. Of International Conference on Multimedia Retrieval ACM ICMR 2018, 11-14 June, Yokohama, Japan, 2018.

[14] Z. Theodosiou, N. Tsapatsoulis, "Image Retrieval Using Keywords: The Machine Learning Perspective, in Semantic Multimedia Analysis and Processing, Ed. By E. Spyrou, D. Iakovides, P. Mylonas, CRC Press \ Taylor & Francis, 2014.

[15] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning, Nature, vol. 521, pp. 436444, 2015.

[16] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. Yuille, "DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 40, pp. 834–848, 2018.

[17] B. Athiwaratkun, K. Kang, "Feature Representation in Convolutional Neural Networks, arXiv preprint arXiv:1507.02313, 2015.

[18] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "CNN features off-the-shelf an astounding baseline for recognition, CoRR, vol. abs\1403.6382, 2014.

[19] A. Penna, S. Mohammadi, N. Jojic, V. Murino, "Summarization and classification of wearable camera streams by learning the distributions over deep features of out-of-sample image sequences, In Proc of IEEE International Conference on Computer Vision (ICCV), pp. 4326–4334, 2017.

[20] Z.Wang, M.D.Hoffman, P.R.Cook, and K.Li, "Vferret:content-based similarity search tool for continuous archived video", In ACM workshop on Continuous archival and retrieval of personal experiences, pp.19–26, 2006.

[21] O. Aghazadeh, J. Sullivan, and S. Carlsson, "Novelty detection from an ego-centric perspective", In proc. of IEEE Conference on Computer Vision and Pattern Recognition, pp. 3297–3304, 2011.

[22] P. Wang and A. F. Smeaton, "Semantics-based selection of everyday concepts in visual lifelogging", International Journal of Multimedia Information Retrieval, vol. 1(2), pp. 87–101, 2012.

[23] W. Min, X. Li, C. Tan, B. Mandal, L. Li, and J.-H. Lim, "Efficient retrieval from large-scale egocentric visual data using a sparse graph representation", In proc. of IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 541–548, 2014.

[24] G. Oliveira-Barra, M. Bolaos, E. Talavera, O. Gelonch, M. Garolera, P. Radeva, "Chapter 7 - Lifelog retrieval for memory stimulation of people with memory impairment", Editor(s): Xavier Alameda-Pineda, Elisa Ricci, Nicu Sebe, In Computer Vision and Pattern Recognition, Multimodal Behavior Analysis in the Wild, Academic Press, pp. 35–158, 2019.

[25] J. Gemmell, G. Bell, and R. Lueder, "MyLifeBits: a personal database for everything", Commun. ACM, vol. 49 (1) (January 2006), pp. 88–95, 2006.

[26] S. Ding, S. Qu, Y. Xi, and S. Wan, "A long video caption generation algorithm for big video data retrieval", Future Generation Computer Systems, vol.93, pp. 583–595, 2019.

[27] L. Zhou, Z. Hinbarji, D.-T. Dang-Nguyen, C. Gurrin, "Lifer: an interactive lifelog retrieval system", In Proc. of ACM Workshop on The Lifelog Search Challenge, LSC 2018, pp. 9-14. ACM, New York, 2018.

[28] B. Münzer, A. Leibetseder, S. Kletz, M. J. Primus, K. Schoeffmann, "lifeXplore at the lifelog search challenge 2018", In Proc. of ACM Workshop on The Lifelog Search Challenge, LSC 2018, pp. 3-8. ACM, New York, 2018.

[29] J. Lokoc, W. Bailer, K. Schoeffmann, B. Muenzer, G. Awad, "On influential trends in interactive video retrieval: video browser showdown 20152017", IEEE Trans. Multimed., vol. 20(12), pp. 3361-3376, 2018.

[30] A. Duane, C. Gurrin, W. Huerst,"Virtual reality lifelog explorer: lifelog search challenge at ACM ICMR 2018", In Proc. of ACM Workshop on The Lifelog Search Challenge, LSC 2018, pp. 20-23. ACM, New York, 2018.

[31] A. Alsina, X. Giró, C. Gurrin, "An interactive lifelog search engine for LSC2018", In Proc. of ACM Workshop on The Lifelog Search Challenge, LSC 2018, pp. 30-32. ACM, New York, 2018.

[32] T.-D. Truong, T. Dinh-Duy, V.-T. Nguyen, M. -T. Tran, "Lifelogging retrieval based on semantic concepts fusion", In Proc. of ACM Workshop on The Lifelog Search Challenge, LSC 2018, pp. 24-29. ACM, New York, 2018.