Towards a Rich and Dynamic Human Digital Memory in Egocentric Dataset

Khalid EL Asnaoui, Mohamed Ouhda, Brahim Aksasse, Mohammed Ouanan
Moulay Ismail University, Faculty of Sciences and Techniques
Morocco
khalid.elasnaoui@gmail.com
ouhda.med@gmail.com
baksasse@yahoo.fr
ouanan_mohammed@yahoo.com

ABSTRACT: *Memories have always been a considerable importance of a persons life and experiences. A digital human memory as a field of study focuses on encapsulating this phenomenon, in digital form, during the thread of a lifetime. By spreading hardware everywhere, massive amount of data is being generated together by people and the surrounding environment. With all this demountable information available, successfully exploring, researching and collating, together, to form a human digital memory, is a new challenge and requires novel and efficient algorithmic solutions. The main goal of this work is going to propose a new method to automatically create rich and dynamic human digital memory in egocentric dataset from the lifelogging images of a person. For this purpose, we will propose a technique using Convolutional Neural Network (CNN) model. For validation, we will apply the proposed method on the Egocentric Dataset of University of Barcelona (EDUB) of 4912 daily images acquired by four persons.*

Keywords: Lifelogging, Visual Lifelogs, Human Digital Memory, CNN, EDUB

Received: 1 June 2017, Revised 29 June 2019, Accepted 6 July 2017

1. Introduction

Images on the Internet and in multimedia systems are rising successively. There are different research works on visual information and automatic analysis of images. Image memorability is a new task in computer vision. Actually, the human brain processes simultaneously millions of images and other information from multiple sources. Among these various images and information some of them are more memorable than others. Studying images memorability is an image processing task, which tends to define and establish the characteristics of memorable images. These characteristics are used to create a representative model for predicting images memorability.

Humans can remember thousands of images and a remarkable amount of their visual details [1]. While, some images are ignored or forgotten quickly [2], others are more memorable, ie. they are engraved in the human memory from the first exposure. Various

fields are concerned with images memorability, for example: the field of advertising (TV, magazine, etc.) and the field of communication (television programs, presentations, etc.) since they seek to attract the attention of viewers to convey their ideas. In order to understand the notion of image memorability, various studies [3] - [5] were carried out along two axes: a psychological / physiological axis because the researchers found a relation between the visual perception of the human being, Its psychic state and images memorability, and a technical axis dealing with the descriptors (color, gradient, texture, semantics, etc.) characterizing these images, the movement of the eye and several other properties. Thus, the authors of [3] carried out a study in order to find memorability scores for a set of images. Thus, a memorability score represents the percentage of correct detections by participants. Visualizations are tools used for creating images, diagrams, or data representations to communicate a message. Visualization through visual imagery is an effective way to communicate both abstract and concrete ideas [6]. Different visual designs offer significantly distinct reading accuracy [7]. While memorability and visualizations memorability remain a new axis of study. The memorability of visualization as an image constitutes an intrinsic property of this visualization [8]. Thus, the visual characteristics extracted from the visualizations can explain their memorabilities. A representative model of visualization memorability can be realized based on a set of features inherent to this visualization.

In addition, Maite Garolera, from the neuropsychology unit of Terrassa Hospital (Spain) said: The brain is designed to forget, we need to forget to survive, because we cant live remembering in each moment all that we have lived. For that reason, humans developed several tools and alternatives for persistent memory, like writings, drawings or photographs. These tools have artificially extended our capabilities for knowledge discovery and transference. Wearable cameras are new tools that take one step further, by allowing a much finer visual memory [9].

First person vision (FPV) can now be captured by the egocentric camera. There are many brands and models of egocentric cameras, with different image quality, lenses but the purpose of these devices is the same, capture peoples life.

Lifelogging consists of taking over a long period of time images that capture the daily experiences and activities of the user wearing a camera. Lifelogging is an active research field for which several devices are spreading faster every day [10].

A lifelogger is a person who captures his daily life to create a virtual and digital memory of his life [11].

Our days have an average of 16 hours awake or 960 minutes. A large amount of data is accumulated in that 960 minutes, on average of 1.400-2.000 images per day, because the lifelogging cameras usually take a photo every 30 seconds. This image storage equals about two gigabytes per day, with most imaging blurred from rapid movements or fast lighting changes. In addition, to this big data problem, most of these images are unintentionally taken, consequently, many of them can be blurred (because a camera wearer is moving quickly), with little information, fast illumination changes. To summarize, wearable cameras define a problem of big, noisy, unlabeled and unstructured visual data that present major challenges and require automatic and efficient algorithmic solutions for finding, indexing and retrieval specific images in this huge amount of data. The rest of the paper is organized as follows: in the section 2, a variety of life-logging wearable devices is presented, section 3 deals with some related work. In section 4, we will describe our contribution. Section 5 presents the data used on which we will apply our method. The conclusion is given in the last section.

2. Life-logging Wearable Devices

The collection of these data tools goes from smartphone to automated digital camera via electronic bracelets (see Fig. 1.(f)) and more generally, the wearable instruments that are carried themselves. The lifelogging is closely related to the movement of the quantified self . The information used to be archived for the benefit of the lifeLogger, and shared with others in various degrees.

The arrival of the smartphone and more recently technologies of wearable technologies really created the opportunity for mass participation in lifelogging, since prior to that, all of the hardware required was very specific or proprietary. Nowadays, wearable cameras are very small, discrete cameras housed as watches, glasses and other subtle wearable devices that can be worn all the day and automatically record a persons everyday activities in a passive fashion. Most wearable cameras in the market such as MeCam, GoPro (see Fig. 1.(a)), Google Glass (see Fig. 1.(c)) or Looxcie (see Fig. 1.(b)) are called wearable video cameras, and have relatively high-temporal resolution (HTR). They capture around 35 frames per seconds and are mostly used for recording the user experience during a few hours in sports and entertainment over the last years [10-11].

Instead, lifelogging photo cameras, such as the Narrative Clip (see Fig. 1.(e)), are prepared to take pictures in a time-lapse mode

that can cover all the day. In this work, we use a Narrative camera that has gone on sale in late 2012, and incorporates a small wearable camera that clips onto the clothes of the wearer to capture over a thousand images per day using the in-built optical sensor. Usually, these cameras upload their images to their corresponding cloud-based server for online display, event segmentation and analysis. SenseCam (see Fig. 1.(d)), that initially created by Microsoft is a wearable camera, worn about the neck that can capture thousands of photos daily, have low-temporal resolution (LTR) and capture only 2-3 frames per minutes, being suitable for image acquisition during long periods of time [10-11].

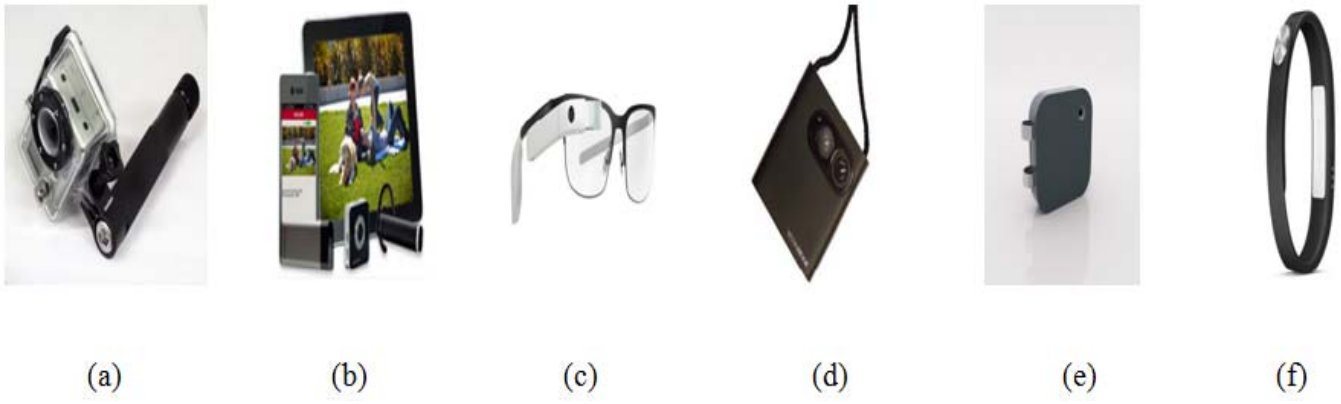

Figure 1. A variety of life-logging wearable devices: (a) GoPro (2002) (b) Looxcie (2011) (c) Google glasses (d) SenseCam (2005) (e) Narrative (2013) (f) Bracelet

The pictures taken provide considerable potential for knowledge how people live their lives, hence, they provide new opportunities for many applications in several fields including security, leisure, healthcare, and the quantified self. The data can be images, texts, sounds, numerical measures such as cardiac rhythm, sleep time or calories ingested and all biological data from sensors on the body and all the moments that matter to you, your workouts to keep fit at your hobbies. So run, walk in the park, browsing the web or watching a movie: you keep track of everything you do. The lifelogs formed by the data collected, over long periods of time, by continuously recording the user life, provide a large potential of mining or inferring knowledge about how people live [1], hence enabling a large amount of applications. Indeed, a collection of studies published in a special issue of the American Journal of Preventive Medicine [12] has proved the potential of visual lifelogs captured through a SenseCam from several viewpoints. In particular, it has been demonstrated that, used as a tool to understand and track lifestyle behavior, visual lifelogs would enable the prevention of non-communicable diseases associated to unhealthy trends and risky profiles (such as obesity or depression, among others). In addition, the lifelogs can be used as a tool for re-memory cognitive training; visual lifelogs would enable the prevention of cognitive and functional decline in elderly people [13]-[15].

When analyzing several days of a person (lifelogger) trying to characterize his behavior, habits or lifestyle, a natural question arises: how to automatically create rich human digital memory. Such information can be much interesting for different health applications: for example, days when the wearer of the camera is less active could predict beginning of depression or physical pain, while days that are too busy could lead to stress and fatigue. Our main goal in this work is going to present new tools by introducing the concept of visual memorability and the contextual data captured from physiological signals. Toward this end, we will develop and test our algorithm to create rich and dynamic human digital memory based on 4912 daily images acquired by four persons using a wearable camera.

3. Related work

In recent years, many interesting applications for lifelogging and human behavior have appeared and are being actively researched. Visual lifelogging data have been used to address different computer vision problems: informative image detection [16-17], egocentric summarization [18-19], content-based search and retrieval [20-21], interaction analysis [22-23], scene understanding [24], concept recognition [25], body movements [26], object-hand recognition [27-28], just to mention a few. Especially interesting is the work on behavior analysis from egocentric data. Fathi et al. [29] presented a new model for human activities recognition in short egocentric videos. This work is extended in [30] by a generative model incorporating the gaze features. Pirsiavash et al. [31] presented a temporal pyramid to encode spatio-temporal features along with detected active

objects knowledge. Moreover, Ma et al. [32] proposed a twin stream Convolutional Neural Network (CNN) architecture for activity recognition from videos.

The memorability of an image is a topic that has been appeared recently in the research community. These works [3-4], [33]-[35] study what makes an image memorable and explore visual features and composition of the image.

Decide when an image is memorable or not, might be interesting for advertising, photography, etc. Generally, memorability is useful because from a day, we want to select those images that are relevant and we will assume that an image is relevant if we can remember it. Indeed, the work presented in [3] defines that an image is memorable if we just saw it for a second, we can detect that it is a repetition. This assumption seems to be coherent. The researchs authors did an experiment with many users. The task for the users was a visual memory game. The game consists in an application that shows images during 5 minutes. They define two types of images: the targets are the images that they want to annotate. There are also fillers, the images that they put between targets. Targets are shown only twice, while fillers can be repeated many times. Some fillers are considered for vigilance, to ensure that a user is paying attention in the game and the results of the game can be used in the research. With the data collected during the game, they compute a memorability score for each image.

Dobbins et al. [36] suggested the DigMem system, which used distributed mobile services, machine learning and linked data in order to create such memories. Along with the design of the system, a prototype had been developed, and two case studies have been undertaken, which successfully created memories.

A DigMem has been presented in [37]. Indeed, it was a platform for creating human digital memories, based on device-specific services and the users current environment. Therefore, information was semantically structured to create temporal memory boxes for human experiences. A working prototype has been successfully developed, which demonstrated this approach.

Borkin et al. in [8], carry out a visualization study using 2,070 single-panel visualizations, categorized with visualization type (e.g., bar chart, line graph, etc.), collected from different sources: news media sites, government reports, scientific journals, and infographics. They assign memorability scores for hundreds of these visualizations, and they find that these score are consistent between observers. Thus, memorability is an intrinsic property of visualizations. Also, they annotate visualization with different attributes like: ratings for data-ink ratios and visual densities. These attributes are used to analyze visualizations memorability.

Kim et al. [38] presented a crowd-sourced study, in which they investigate the utility of using mouse clicks as an alternative for eye fixations in the context of understanding data visualizations. Participants were presented with a series of images containing graphs and diagrams and asked to describe them. Each image was blurred so that the participant needed to click to reveal bubbles - small, circular areas of the image at normal resolution. Then, they compare the bubble click data with the fixation data from a complementary eye-tracking experiment by calculating the similarity between the resulting heat maps. Thus, a high similarity score suggests that the proposed methodology is a practical crowd-sourced alternative to eye tracking experiments. Actually, they design their approach to measure which information people consciously choose to examine for understanding visualizations.

The work presented in [39], examines how visualizations are recognized and recalled. They annotate a dataset of 393 visualizations and analyze the eye movements of 33 participants. Also, they gather text descriptions of the visualizations generated by thousand of participant. Thus, they determine what components of visualizations attract peoples attention, and what information is encoded into memory.

Following the same objective of human digital memory, the question is how to use visual lifelogs for it, since lifelogging images give much richer information about human behavior since visual lifelogging images contain information about the environment of the person, the events he/she is involved, interactions and daily activities, etc. To the best of our knowledge, this is the first work suggesting an automatically creating rich human digital memory using lifelogging images.

4. Description of the Proposed Contribution

As we mentioned, memories are an important aspect of a persons life and experiences. The digital domain of human memories focuses on the encapsulation of this phenomenon, in digital form, during the thread of a lifetime. By spreading hardware everywhere, people and their environments generate an enormous collection of data. With all of this demountable information

available, successfully exploring, researching and collating, to form a human digital memory, is a challenge. This is especially true when the age of the data is verified. The linked data provides an ideal and new solution to overcome this challenge, where a variety of their sources can be extracted for detailed information about a given event. The digital human memory, created with the data from the lifelogging devices, produces a dynamic and rich memory. Memories, created in this way, contain living structures and various data sources, which result from the semantic compilation of content and other memories. Information can be created as how you feel, where you are, and the context of the environment.

The contribution of this paper consists to introduce the concept of visual lifelog to create a dynamic and rich human digital memory using visual image content. In this section a method to compute memorability maps will be presented. A memorability map is an image where each pixel has a normalized value between 0 and 1 related with the contribution of this region to the visual memorability of the image. By this way, we can know what parts or regions of an image make it memorable.

Furthermore, there is a map called saliency map that describe with an image where the human fix his eye gaze where he observes a scene in the firsts moments. From memorability and saliency maps, it is interesting to relate both maps to know what parts of the image make it memorable. Toward this goal, we need to compute saliency and memorability maps for each image. Saliency maps have been explored before and there is one convolutional neural network model that obtains these maps from image. SalNet is a CNN model for saliency map prediction that will be used due to the good results obtained in a challenge for this purpose. The result of the algorithm is a grey map where saliency points are labeled with a value near to 1 while points out of the eye gaze have a label near 0.

To our best knowledge, there is no algorithm or CNN model to compute memorability maps using EDUB. For this reason, our contribution will be the first one that will use CNN model to create dynamic and rich human digital memory using EDUB [40].

5. Experimental Results

Our work is aimed at creating rich human digital memory, usually captured with lifelogging wearable camera.

5.1 Data

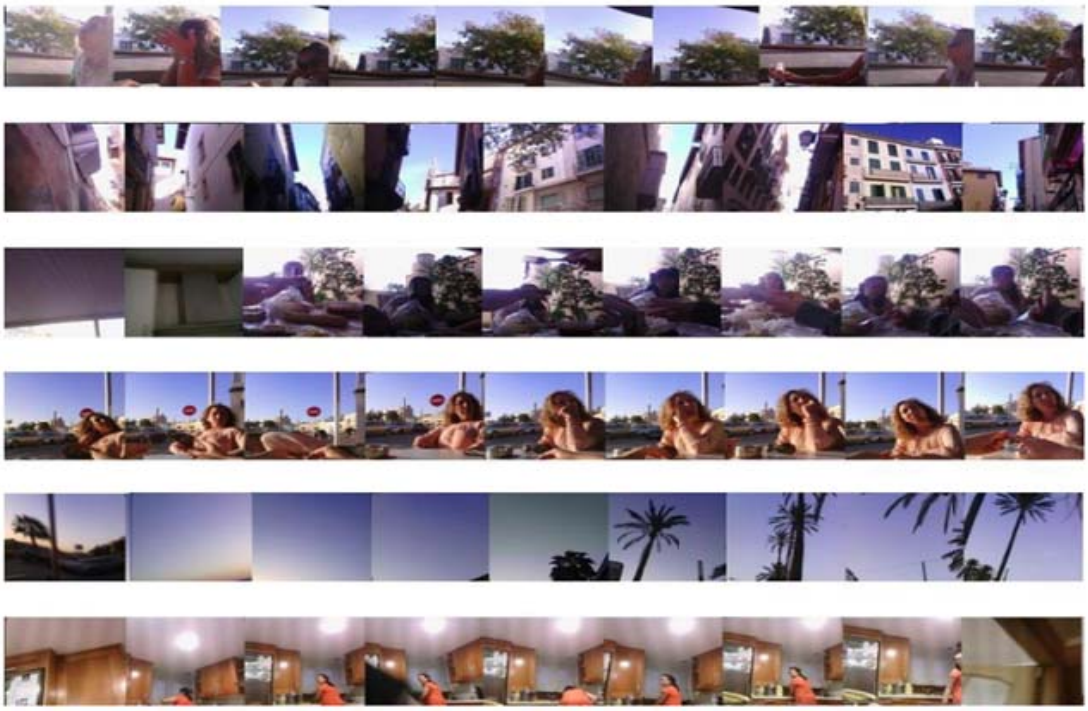

Figure 2. Example of images of the EDUB dataset acquired.

For this purpose, our experiments will be performed on the EDUB public dataset of images acquired with a Narrative wearable camera. This device is typically clipped around the chest area or on the users clothes under the neck. For this purpose, we will use the public image dataset EDUB http://www.ub.edu/cvub/dataset/ on which we will validate our results (Figure 2). This dataset is a set composed of 4912 images, their sizes are 384x512, and is acquired by four persons using Narrative camera (see Fig. 1.(e)). This set is acquired during 8 different days, two days per person. Figure 2 shows an example of the images in the EDUB dataset [40].

6. Conclusion

In this paper, we have addressed the following problem: how to create a rich human digital memory captured by a wearable camera. We presented new approach that will use convolutional neural networks which have the capability to keep localization in images. This capability is due to the convolutional layers because there are built with convolutional filters that have local impact. The proposed approach will be able to automatically create rich human digital memory. Although the work presented a preliminary validation, we believe it demonstrates the potential of lifelogging techniques to create dynamic and rich memory.

References

[1] Standing. L. (1972). Learning 10,000 pictures. *Q J Exp Psychol*, 25, p. 207-222.

[2] Konkle, T., Brady, T. F., Alvarez, G. A., Oliva. A. (2010). Scene memory is more detailed than you think: the role of categories in visual long term memory. *Psychological Science*, 21 (11) 1551-1556.

[3] Isola., P., Xiao., J., Torralba, A., Oliva, A. (2011). What makes an image memorable? *In:* Computer Vision and Pattern Recognition (CVPR), *IEEE Conference on*. 145-152.

[4] Isola, P., Parikh, D., Torralba, A., Oliva, A. (2011). Understanding the intrinsic memorability of images. *In*: Advances in Neural *Information Processing Systems*. p. 2429-2437.

[5] Mancas, M., Le Meur, O. (2013). Memorability of natural scene: The role of attention. ICIP, p. 196-200.

[6] Cairo, A. (2013). The Functional Art: An Introduction to Information Graphics and Visualization, New Riders.

[7] Cleveland, S. W., MCGILL, R. (1984). Graphical perception: Theory, experimentation, and application to the development of graphical methods. *J. American Statistical Association*, 79, p. 531-554.

[8] Borkin, M. A., Vo, A. A., Bylinskii, Z., Isola, P., Sunkavalli, S., Oliva, A., Pfister, H. (2013). What makes a visualization memorable?. *IEEE TVCG*, 19 (12) 2306-2315.

[9] Herrera, M. C. (2016). Visual Memorability for Egocentric Cameras. Universitat Politcnica de Catalunya Escola Superior denginyeria Industrial, Aeroespacial I Audiovisual de Terrassa.

[10] Asnaoui, EL., K., Petia, R., Aksasse, B., Ouanan. M. (2017). Using Content-Based Image Retrieval to automatically assess day similarity in visual lifelogs. Selected as best paper in The International Conference on Intelligent Systems and Computer Vision. *IEEE Conference Publications*.

[11] Asnaoui, EL., K., Aksasse, H., Aksasse, B., Ouanan, M. (2017). A Survey of Activity Recognition in Egocentric Life-logging datasets, *In:* International Conference on Wireless Technologies, embedded and intelligent Systems. *IEEE Conference Publications*.

[12] Doherty, A. R., Hodges, E. S., King, A. C., Smeaton, A. F., Berry, E., Moulin, J C. Lindley, An., Kelly, Paul., Foster, Charlie (2013). Wearable cameras in health: the state of the art and future possibilities. *American Journal of Preventive Medicine*, 44 (3) 320 - 323.

[13] Hodges, S., Williams, L., Berry, E., Izadi, S., Srinivasan, J., Butler, A., Smyth, G., Kapur, N., Wood, K. (2006). Sensecam: A retrospective memory aid. *In:* UbiComp: *Ubiquitous Computing*, p. 177 -193. Springer.

[14] Doherty, A. R., Pauly-Takacs, K., Caprani, N., Gurrin, C., Moulin, JA., C., OConnor, N. E., Smeaton, A. F. (2012). Experiences of aiding autobiographical memory using the sensecam. *Human Computer Interaction*, 27 (1-2) 151- 174.

[15] Lee, M. L., Dey, A. K. (2008). Lifelogging memory appliance for people with episodic memory impairment. *In*: Proceedings of the 10th International Conference on Ubiquitous Computing, p. 44 - 53. ACM.

[16] Xiong, B., Grauman, K. (2014). Detecting snap points in egocentric video with a web photo prior. *In*: Computer Vision,European Conference on, p. 282298. Springer.

[17] Lidon, A., Bolaos, M., Dimiccoli, M., Radeva, P., Garolera, M., Gir-i Nieto, X. (2015). Semantic summarization of egocentric photo stream events.

[18] Smeaton, A. F., Over, P., Doherty, A. R. (2010). Video shot boundary detection: Seven years of trecvid activity. *Computer Vision and Image Understanding*, 114 (4) 411-418.

[19] Jinda-Apiraksa, A., Machajdik, J., Sablatnig, R. (2012). A keyframe selection of lifelog image sequences. Erasmus Mundus M. Sc. *In:* Visions and Robotics thesis, *Vienna University of Technology*.

[20] Wang, Z., Hoffman, M. D., Cook, P. R., Li. K. (2006). Vferret: content-based similarity search tool for continuous archived video. *In*: ACM workshop on Continuous archival and retrieval of personal experiences, p. 1926.

[21] Chandrasekhar, V., Tan, C., Min, W., Liyuan, L., Xiaoli, L., Hwee, J. L. (2014). Incremental graph clustering for efficient retrieval from streaming egocentric video data. In Pattern Recognition, *IEEE International Conference* on, p. 2631 - 2636.

[22] Doherty, A. R., Smeaton, A. F. (2008). Combining face detection and novelty to identify important events in a visual lifelog. *In:* Computer and Information Technology Workshops, IEEE International Conference on, p 348 - 353.

[23] Alletto, S., Serra, G., Calderara, S., Cucchiara, R. (2014). Head pose estimation in firstperson camera views. *In:* Pattern Recognition (ICPR), 22nd International Conference on. IEEE.

[24] Kikhia, B., Boytsov, A.Y., Hallberg, J., Jonsson, H., Synnes, K. (2014). Structuring and presenting lifelogs based on location data. *In:* Pervasive Computing Paradigms for Mental Health, p 133-144. Springer.

[25] Byrne, D., Doherty, A. R., Snoek, C. G. M., Jones, G. J. F., Smeaton, A. F. (2010). Everyday concept detection in visual lifelogs: validation, relationships and trends. *Multimedia Tools and Applications*, 49 (1) 119-144.

[26] Kitani, K. M., Okabe, T., Sato, Y., Sugimoto, A. (2011). Fast unsupervised ego-action learning for first-person sports videos. *In:* Computer Vision and Pattern Recognition, IEEE Conference on, p 3241-3248.

[27] Fathi, A., Farhadi, A., Rehg, J. M. (2011). Understanding egocentric activities. In Computer Vision, *IEEE International Conference on*, p. 407-414.

[28] Sundaram, S., Mayol-Cuevas, W. W. (2010). Egocentric visual event classification with location-based priors. *In*: Advances in Visual Computing, p. 596605. Springer.

[29] Fathi, A., Farhadi., Rehg, J. M. (2011). Understanding egocentric activities. *In*: 2011 International Conference on Computer Vision, p. 407-414. IEEE.

[30] Fathi, A., Li, Y., Rehg, J. M. (2012). Learning to recognize daily actions using gaze. *In*: European Conference on Computer Vision, p. 314-327. Springer.

[31] Pirsiavash, H., Ramanan, D. (2014). Parsing videos of actions with segmental grammars. *In*: Computer Vision and Pattern Recognition (CVPR).

[32] Ma, M., Fan, H., Kitani, K. M. (2016). Going deeper into first-person activity recognition. *In*: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

[33] Khosla, A., Xiao, J., Isola, P., Torralba, A., Oliva, A. (2012). Image memorability and visual inception. *In*: SIGGRAPH Asia 2012 Technical Briefs. p. 35. ACM.

[34] Khosla, A., Xiao, J., Torralba, A., Oliva, A. (2012). Memorability of Image Regions. *In*: NIPS, Vol. 2, p. 4.

[35] Bainbridge, W. A., Isola, P., Oliva, A. (2013). The intrinsic memorability of face photographs. *Journal of Experimental Psychology: General*, 142 (4) 1323.

[36] Dobbins, C., Merabti, M., Fergus, P., Jones, D. L. (2014). Creating human digital memories with the aid of pervasive mobile devices. *Pervasive and Mobile Computing* 12, p. 160178.

[37] Dobbins, C., Merabti, M., Fergus, P., Jones, D. L., Bouhafs, F. (2013). Exploiting linked data to create rich human digital memories. *Computer Communications* 36, p. 16391656.

[38] Borkin, M., Bylinskii, Z., Kim, N., Bainbridge, C., Yeh, C., Borkin, D., Pfister, H., Oliva, A. (2016). Beyond memorability:

Visualization recognition and recall. *IEEE Transactions on Visualization and Computer Graphics*, 22(1) 519-528.

[39] Kim, N., Bylinskii, Z., Borkin., M., Oliva, A., Gajos, K. Z., Pfister, H. (2015). A Crowd sourced Alternative to Eye-tracking for Visualization Understanding. CHI15 Extended Abstracts. Seoul, Korea: ACM, 1349-1354.

[40] Bolaos, M., Dimiccoli, M., Radeva, P. (2015). Towards Storytelling from Visual Lifelogging: An Overview. *Journal of Transactions on Human-Machine System*s.