Image Aesthetics and Content in Selecting Memorable Keyframes from Lifelogs

Feiyan Hu and Alan F. Smeaton✉

Insight Centre for Data Analytics
Dublin City University, Dublin 9, Ireland
{alan.smeaton}@dcu.ie

Abstract. Visual lifelogging using wearable cameras accumulates large amounts of image data. To make them useful they are typically structured into events corresponding to episodes which occur during the wearer's day. These events can be represented as a visual storyboard, a collection of chronologically ordered images which summarise the day's happenings. In previous work, little attention has been paid to how to select the representative keyframes for a lifelogged event, apart from the fact that the image should be of good quality in terms of absence of blurring, motion artifacts, etc. In this paper we look at image aesthetics as a characteristic of wearable camera images. We show how this can be used in combination with content analysis and temporal offsets, to offer new ways for automatically selecting wearable camera keyframes. In this paper we implement several variations of the keyframe selection method and illustrate how it works using a publicly-available lifelog dataset.

Keywords: Lifelogging, keyframes, image aesthetics, image quality

1 Introduction to Lifelogging

Lifelogging is a phenomenon of automatically and ambiently recording different aspects of ordinary, everyday life, in digital format [7]. This has become a topic of research interest and practical use because of the development of wearable sensors and their reduction in size and most importantly the way battery technology has improved to the point of enabling all-day continuous recording.

Lifelogs can be generated using a range of wearable sensors including physiology sensors (heart rate, respiration, etc.), activity sensors (wrist-worn accelerometers), location sensors (GPS and indoor location tracking), environmental sensors (passive infra-red for detecting presence, temperature, humidity, etc.) and wearable cameras which record what the user is doing and experiencing, from the wearers' viewpoint. The most popular wearable cameras are worn on the chest, are front-facing and have a wide-angle lens to record a broad perspective of the viewers point of view [6]. Many devices like the GoPro and similar can record continuous HD video, as well as audio. For niche applications like wearable cameras for law enforcement, this is acceptable but leads to storage requirements which are excessive for scalable lifelogging and for less specialist uses.

The most popular wearable camera devices used for lifelogging is the Autographer and prior to that it was the Narrative and before that the SenseCam. Functionally these are all quite similar in that they each take several thousands of images per day, usually triggered by on-board sensors such as an accelerometer to detect movement. In general these take about 2 or 3 images per minute and store these on-board for later downloading and processing. The processing usually involves structuring a lifelog into discrete and non-overlapping events and selecting single image keyframes from each event as representative of the activity in the event [4].

The selection of the keyframe to use as the event summary has not been a subject of much investigation and simple techniques such as choosing the keyframe in the middle of the event, or the first or last, or the one with best image quality, have generally been used. In this paper we re-examine the question of "which lifelog image to use to summarise an event" by exploring different aspects of lifelog images including image quality, image content and image aesthetics, as well as combinations of them. We present results computed from a publicly available lifelog dataset which compares different approaches.

2 Keyframes from Visual Lifelogs

There are many use cases for lifelogging including self-monitoring of sleep or activity levels for health and wellness, long term monitoring for supporting behaviour change like smoking cessation, activity recording by personel in security settings, activity and event recording in certain employment areas like health professionals [7]. The application that we are interested in is memory augmentation and memory support helping them to remember and to learn better and to remember more and to remember things that are more important. While we currently focus on people without memory impairment, ultimately this can have possibilities for people with various forms of dementia as shown in the preliminary work by Piasek *et al.* [16].

Harvey *et al.* [8] have argued that the increasing interest in and development of lifelogging does present clear possibilities for using technology, specifically technology which generates lifelogs, to augment human memory beyond what is currently done, which is mostly just about reminders and prompts. Their work does note the ethical concerns and dangers with doing this and that we should be aware of moving beyond prompts and reminders and into augmentation. Lets not forget that there are reasons why sometimes we do want to forget. Silva *et al* go further in [18] and point to a lack of theory behind memory augmentation which can guide us on how to use visual lifelogging in memory augmentation or rehabilitation. Most of the studies to date have been small in scale and in sample size and evaluation of the efficacy of any form of memory augmentation has always been difficult.

The basic premise on which almost all (visual) lifelog applications are based, especially those which address memory rehabilitation or support, is to present a visual summary of each day as a storyboard of images, a selection of images taken

from the wearable camera. These are usually filtered to eliminate poor quality images, including those with blurring or occlusion caused by hands similar to those shown in Figure 1. However once the poor quality images are removed there is then little guidance on which images to select. Keyframe selection from lifelogs is different to keyframe selection from video shots when the genre is movies, TV programs, news, or any kind of post-produced material where the shot is structured. In image lifelogs, as in many videos on social media, the shots/events are not as structured and the important things can happen serendipitously

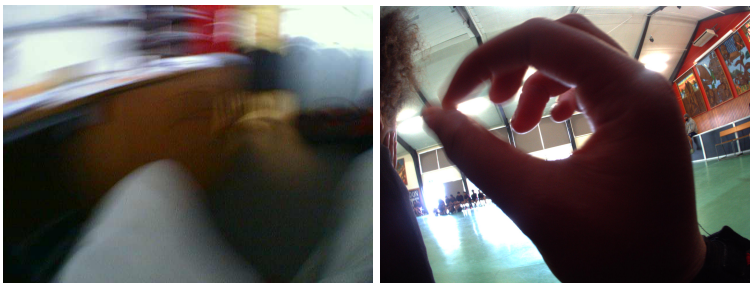

Fig. 1: Examples of poor quality wearable camera images due to wearer movement and occlusion from the wearer's hands, respectively

Doherty *et al.* in [3] did work on automatic selection of keyframe images based on their visual uniqueness, so effectively presenting the day as an illustration of the wearers' most unusual activities, as shown in Figure 2 below. In this rendering of a summary of the day's activities the size of the image is proportional to the visual uniqueness of the image, where uniqueness corresponds to unusual activities for the wearer. This addresses a use case for lifelogging where we want the summary to present unusual events but that's not the same as memory augmentation which is what we are interested in here.

In an early study into the management and use of digital photographs, [17] found that people are more attracted to highly aesthetically attractive pictures. In an even earlier study in [1] it was found that users tend to pick the most aesthetically appealing pictures for their portfolios when asked to choose images of themselves for authentication purposes. More recent work [11] studied the impact that aesthetic images have on people's recollection of news items associated with those images and determined that aesthetics of those associated images does have a big impact on people's views on those stories.

There are many examples in society where people are presented with the task of creating an image that a viewer will remember. That is the basis behind advertising, for example. While it may seem that image memorability is a subjective aspect, not all images are equal in their memorability as shown in [10]. Some will stick in our minds, while others are easily forgotten. It has been shown that image memorability is a characteristic of images that is con-

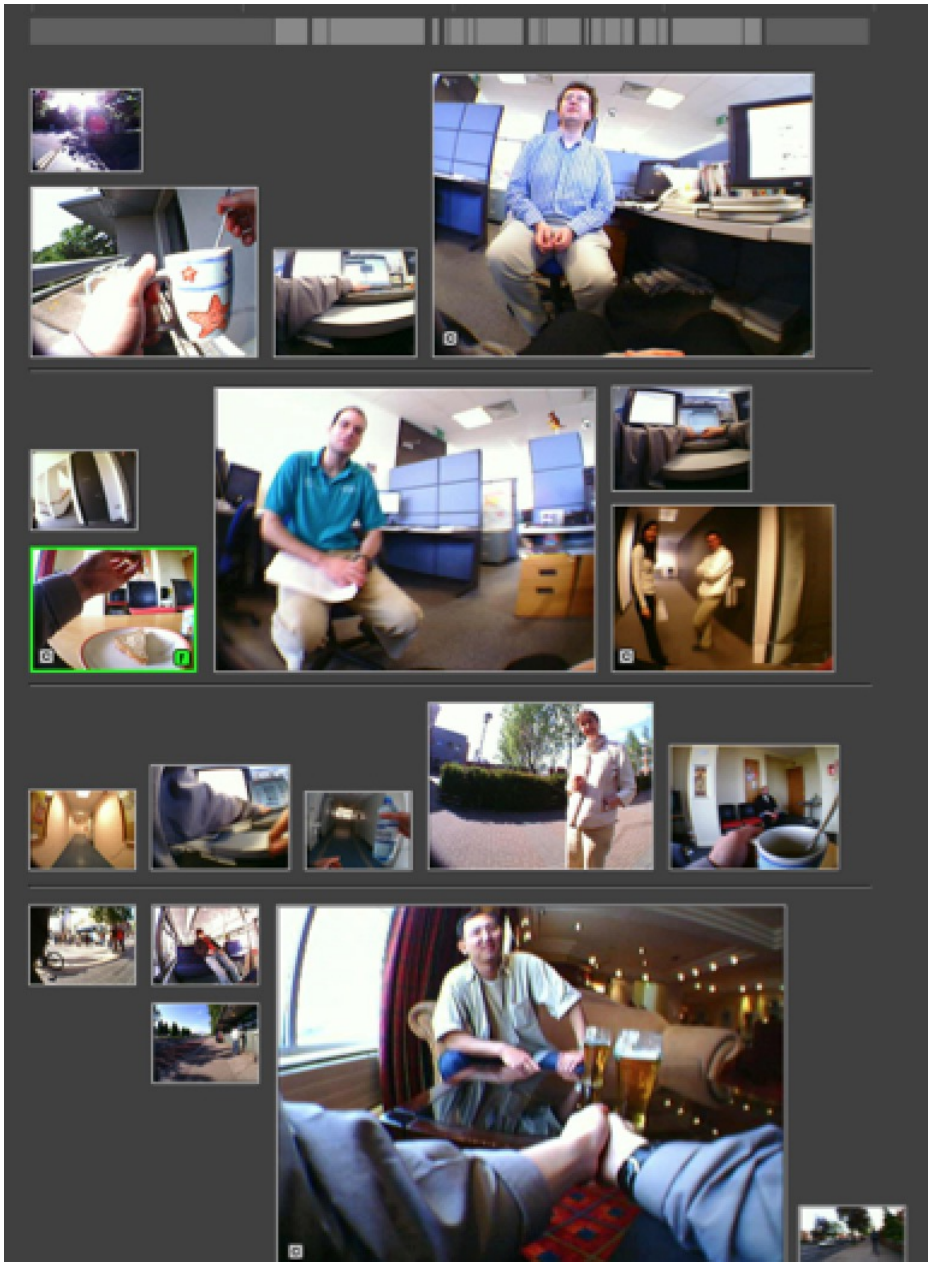

Fig. 2: Doherty's SenseCam Browser, a visual summary of a day taken from [3]

stant and is shared across viewers, in other words different people associate with the same memorability aspects of many images [12, 9]. Given that this is the case, and that our ultimate use case here is triggering memory recall, especially for people with memory impairment, this gives us the rationale for looking at whether we should select the most aesthetically pleasing images from a visual lifelog as summaries of a day. This forms the main criterion for our lifelog event summarisation, computable aesthetics as a proxy for memorability of an image.

3 Computing Image Aesthetics and Uniqueness of Image Semantics

In order to test our ideas on lifelog keyframes we need a lifelog collection which is freely available to allow reprodcibility of our work. Creating and releasing a lifelog collection for public use is one of the most difficult datasets to assemble because of concerns about privacy, ownership of the data. Fortunately such a collection has recently become available.

The NTCIR-13 Lifelog data consists of 5 days of data from two active lifeloggers. The dataset contains biometric data including heart rate, GSR, caloric expenditure, blood pressure, and more, activity data including locations visited, accelerometer data and mood, computer usage data, and the part of interest to us, images taken from a Narrative Clip 2 wearable camera [5]. This is set to take an image at 45 second intervals, corresponding to about 1,500 images per day. With these images there is the accompanying output from an automatic concept detector.

Aesthetics is a fairly ephemeral concept and has to do with the beauty and human appreciation of an object, or whatever is in the image. It is difficult to pin down precisely as it has a subjective element where one person can view a picture or an object as beautiful and another person can have the opposite view. So even though there is no universal agreement or even a ranking of aesthetic quality, and there would be debate about things in the "middle" there's fair enough agreement of things that are, and are not, aesthetically pleasing.

Many computer vision papers have tried to quantify and measure the aesthetic quality of images [13, 14, 2]. Yet this aspect of an image is subjectively derived and aesthetic values of an image will vary from subject to subject. There are some features like sky illumination or certain concepts that have been reported in [2] to have influence on aesthetic scores. With increasing computational power and especially neural networks with pre-trained models, it is now possible to predict or compute aesthetic values for an image. Mai *et al.* [14] used pre-trained models to extract the probability of certain semantic concepts occurring in highly-aesthetic images. Along with probability of concepts, neural networks have also been trained from scratch to compute aesthetics with adaptive pooling layers where combined high level and low level features are used to predict aesthetic scores.

To describe the problem formally we assume that each day a camera captures T images, and each image is I_t where $t = 0...T$. In order to quantify aesthetic

scores, we trained a deep neural network. The network we used is ResNet, pre-trained on ImageNet images to extract image representations and on top of the image representation we add a fully connected layer to predict aesthetic scores. The dataset used to train aesthetic net is from the DPChallenge[1]. The aesthetic score is defined as S_A:

$$S_t^A = f_{NN}(I_t(x, y, c)) \tag{1}$$

where f_{NN} is the trained neural net, and $I(x, y, c)$ is the input image with color channels. Some example lifelog images with their aesthetic scores are shown in Figure 3.

(a) Score: 0.08 (b) Score: 0.11 (c) Score: 0.22 (d) Score: 0.23

(e) Score: 0.54 (f) Score: 0.65 (g) Score: 0.73 (h) Score: 0.75

Fig. 3: Examples lifelog images with their aesthetic scores

In order to determine the uniqueness of each lifelog images in terms of its content, which is a contributing factor to memorability, we use object annotation associated with each image. In the NTCIR Lifelog task, which was described earlier, each image has a number of semantic concepts or objects labeled automatically, and we use $\{O_t\}$ to represent the set of semantics for image t. The number of semantic concepts in each image is defined as:

$$S_t^L = |O_t| \tag{2}$$

[1] `http://www.dpchallenge.com/`

In order to define the uniqueness of a image we define a matrix A_{ij}:

$$A_{ij} = \begin{cases} \frac{S_i^L - |O_i \cap O_j|}{S_i^L} & i \neq j \text{ and } S_i^L \neq 0 \\ 0 & i = j \text{ or } S_i^L = 0 \end{cases} \tag{3}$$

The uniqueness score is then computed as:

$$S_t^U = \sum_{j=0}^{T} A_{tj} \tag{4}$$

The scores are normalized by the maximum score within a day to eliminate inter-daily bias:

$$\begin{aligned} \hat{S_t^A} &= \frac{S_t^A}{\max_t S_t^A} \\ \hat{S_t^L} &= \frac{S_t^L}{\max_t S_t^L} \\ \hat{S_t^U} &= \frac{S_t^U}{\max_t S_t^U} \end{aligned} \tag{5}$$

The process to select key frames of each day of lifelog images is described as:

1. Find the highest n images ranked by aesthetic score S_t^A. This set is marked as $\{\mathcal{A}\}$. In our experiment $n = 100$
2. Find the highest m images ranked by uniqueness of image semantics S_t^U. This set is marked as $\{\mathcal{U}\}$. In our experiment $m = 100$
3. The intersection of $\{\mathcal{A}\}$ and $\{\mathcal{U}\}$ is our candidate set of keyframes $\{\mathcal{K}\} = \mathcal{A} \cap \mathcal{U}$.
4. Images in $\{\mathcal{K}\}$ are ranked in chronological order. Among those ordered images, the time interval between neighboring images less than time s is classified into one group or segment. In our experiment time s is set to 15 minutes.
5. We then select one keyframe from each segment according to different scores or combinations of scores S_t. Different hypothesis to compute S_t are used and these are described in the next section, along with illustrating examples.

4 Creating Storyboards from Lifelog Images

We combined uniqueness of content as represented by concept annotations, image aesthetics and image richness to select keyframes to make storyboards for single days in the NTCIR Lifelog collection. We choose one day from the collection, September 25th, and illustrate the different selection methods for that day, though we would like to have completed a fuller evaluation, which we will return to later.

1. The first method is called **Aesthetics:** and is formally defined as $S_t = \hat{S_t^A}$. The examples of it is shown in Figure 4(a) which shows the timeline as a bar in the middle with the chosen keyframes appearing above and below, and pointing to the time of day when they were taken. There is no supplementary information in this storyboard, just the images and time taken.
2. The second method is called **Uniqueness of semantics:**, formally defined as $S_t = \hat{S_t^U}$ and shown in Figure 4(b). Once again we have a timeline and associated with each image we have the set of annotations assigned to each image. Some of these images, for example the first one of the night sky, may be semantically meaningful but they are not pleasing to look at.
3. The third method is a **Combination of semantic uniqueness and richness:**, defined as $S_t = \frac{1}{2}\hat{S_t^U} + \frac{1}{2}\hat{S_t^L}$ and shown in Figure 4(c) which once again associates semantic concepts or tags with images and also yields a set of images which are at least more pleasing to the eye.
4. In the fourth example we use a **Combination of aesthetic and semantic uniqueness:** which is defined as $S_t = \frac{1}{2}\hat{S_t^A} + \frac{1}{2}\hat{S_t^U}$ and the example is shown in Figure 5(a). There are no concepts to illustrate in this example.
5. The final algorithm to generate storyboard keyframes is called **Combination of aesthetic, semantic uniqueness and richness:** and is defined as $S_t = \frac{1}{2}\hat{S_t^A} + \frac{1}{2}(\frac{1}{2}\hat{S_t^U} + \frac{1}{2}\hat{S_t^L})$. A worked example can be seen in Figure 5(b).

If we look at Figure 4(a), in which keyframes are selected only by aesthetic scores, we notice that even though the third image above the timeline from the left above is considered aesthetically pleasing by the classifier, it doesn't provide much information except that it is an indoor wall. Interestingly when we choose keyframes by combining aesthetic and semantic uniqueness as shown in Figure 5(a), the images chosen at the very same time seem to have much more information. We can tell this event is on a street and can even see the names of some shops. Figure 4(b) shows the storyboard result when using only semantic uniqueness, and it can be observed that most of the selected keyframes are different from those selected by aesthetic value though there are still 3 images that are overlapping, including two images with a laptop screen.

Selection by semantic uniqueness in Figure 4(b) is sensitive to the successful performance of concept detection. A good example to illustrate this is the first image above the timeline on the left. The concepts are mis-classified as *night* and *sky*, which happen to be unique among all the semantics because the wearer did not spend much time outdoors at night. By using the number of concepts appearing in each image it seems it can have leverage on this dilemma. Figure 4(c) seems to return more reasonable results than just using semantic uniqueness alone. The result of aesthetic, semantic uniqueness and richness combined are shown in Figure 5(b). Among the results when using different methods, there are some images that seem to appear repeatedly and have some invariant property. In future work we could extract and further analyse those images.

While the above might seem like a cursory examination of the outputs of different keyframe strategies, a full and thorough evaluation of the *memorability* of the camera images generated by different, and combined, approaches is

out of scope. This would require multiple wearers to generate lifelog content and for each wearer, generate storyboards of their days via all the algorithmic variations mentioned above. We would then present memory recollection tasks to each wearer, for each method, in order to test the efficacy of the different keyframe selection approaches used to generate the storyboards. Such an experiment would need to insulate against the very many confounding variables like wearer variation, time variation, and would make this a huge user experiment. We don't have resources for that so we are limited to observational analysis of generated storyboards presented above.

5 Conclusions

Computing lifelog keyframe selections as described in this paper is not computationally expensive since the aesthetics classifier is already trained and built and all that is required is processing to extract low level features and then run it through the classifier. The early layers of the deep learning network used to compute aesthetics can be re-used as the layers used to extract features for semantic concept recognition and in fact that is what we do when we re-use the layers trained on ResNet and the ImageNet image dataset. So in total, once the training is done this is very fast to run.

There are two main directions we would like to pursue as future work. The first, and most obvious, is a thorough evaluation but we need to develop an evaluation which is not full-on with lots of users involved as sketched out in the previous section since that is neither scalable nor affordable. The second direction is to examine each image for use as a keyframe but not the whole image. Wearable camera images have a wide angle view and they do not capture what the wearer was actually looking at, just what the range of things they may have looked at. Using prior work in saliency detection such as that described in [15], we can identify "parts" or regions within a keyframe which can be a crop from the whole image and then go into the storyboard, rather than the whole image. This is interesting for the memorability application because it can be objects or features within our perspective which trigger memories and this is what that saliency-based cropping yields.

References

1. R. Dhamija and A. Perrig. Deja-Vu a user study: Using images for authentication. In *USENIX Security Symposium*, volume 9, pages 4–4, 2000.
2. S. Dhar, V. Ordonez, and T. L. Berg. High level describable attributes for predicting aesthetics and interestingness. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1657–1664. IEEE, 2011.
3. A. R. Doherty, C. J. A. Moulin, and A. F. Smeaton. Automatically assisting human memory: A SenseCam browser. *Memory*, 19(7):785–795, 2011.
4. A. R. Doherty and A. F. Smeaton. Automatically segmenting lifelog data into events. In *2008 Ninth International Workshop on Image Analysis for Multimedia Interactive Services*, pages 20–23, May 2008.

5. C. Gurrin, H. Joho, F. Hopfgartner, L. Zhou, and R. Albatal. NTCIR Lifelog: The First Test Collection for Lifelog Research. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '16, pages 705–708, New York, NY, USA, 2016. ACM.
6. C. Gurrin, A. F. Smeaton, D. Byrne, N. O'Hare, G. J. F. Jones, and N. O'Connor. An examination of a large visual lifelog. In *Information Retrieval Technology: 4th Asia Information Retrieval Symposium, AIRS 2008, Harbin, China, January 15-18, 2008 Revised Selected Papers*, pages 537–542, Berlin, Heidelberg, 2008. Springer Berlin Heidelberg.
7. C. Gurrin, A. F. Smeaton, and A. R. Doherty. Lifelogging: Personal big data. *Found. Trends Inf. Retr.*, 8(1):1–125, June 2014.
8. M. Harvey, M. Langheinrich, and G. Ward. Remembering through lifelogging: A survey of human memory augmentation. *Pervasive and Mobile Computing*, 27:14–26, 2016.
9. P. Isola, D. Parikh, A. Torralba, and A. Oliva. Understanding the intrinsic memorability of images. In J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, pages 2429–2437. Curran Associates, Inc., 2011.
10. P. Isola, J. Xiao, D. Parikh, A. Torralba, and A. Oliva. What makes a photograph memorable? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1469–1482, 2014.
11. J. Kätsyri, N. Ravaja, and M. Salminen. Aesthetic images modulate emotional responses to reading news messages on a small screen: A psychophysiological investigation. *International Journal of Human-Computer Studies*, 70(1):72–87, 2012.
12. A. Khosla, J. Xiao, P. Isola, A. Torralba, and A. Oliva. Image memorability and visual inception. In *SIGGRAPH Asia 2012 Technical Briefs*, SA '12, pages 35:1–35:4, New York, NY, USA, 2012. ACM.
13. X. Lu, Z. Lin, H. Jin, J. Yang, and J. Z. Wang. Rating image aesthetics using deep learning. *IEEE Transactions on Multimedia*, 17(11):2021–2034, 2015.
14. L. Mai, H. Jin, and F. Liu. Composition-preserving deep photo aesthetics assessment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 497–506, 2016.
15. J. Pan, E. Sayrol, X. Giro-i Nieto, K. McGuinness, and N. E. O'Connor. Shallow and deep convolutional networks for saliency prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 598–606, 2016.
16. P. Piasek, K. Irving, and A. F. Smeaton. SenseCam intervention based on Cognitive Stimulation Therapy framework for early-stage dementia. In *2011 5th International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth) and Workshops*, pages 522–525, May 2011.
17. K. Rodden and K. R. Wood. How do people manage their digital photographs? In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '03, pages 409–416, New York, NY, USA, 2003. ACM.
18. A. R. Silva, M. S. Pinho, L. Macedo, and C. J. A. Moulin. A critical review of the effects of wearable cameras on memory. *Neuropsychological Rehabilitation*, 26(1):1–25, 2016. PMID: 26732623.

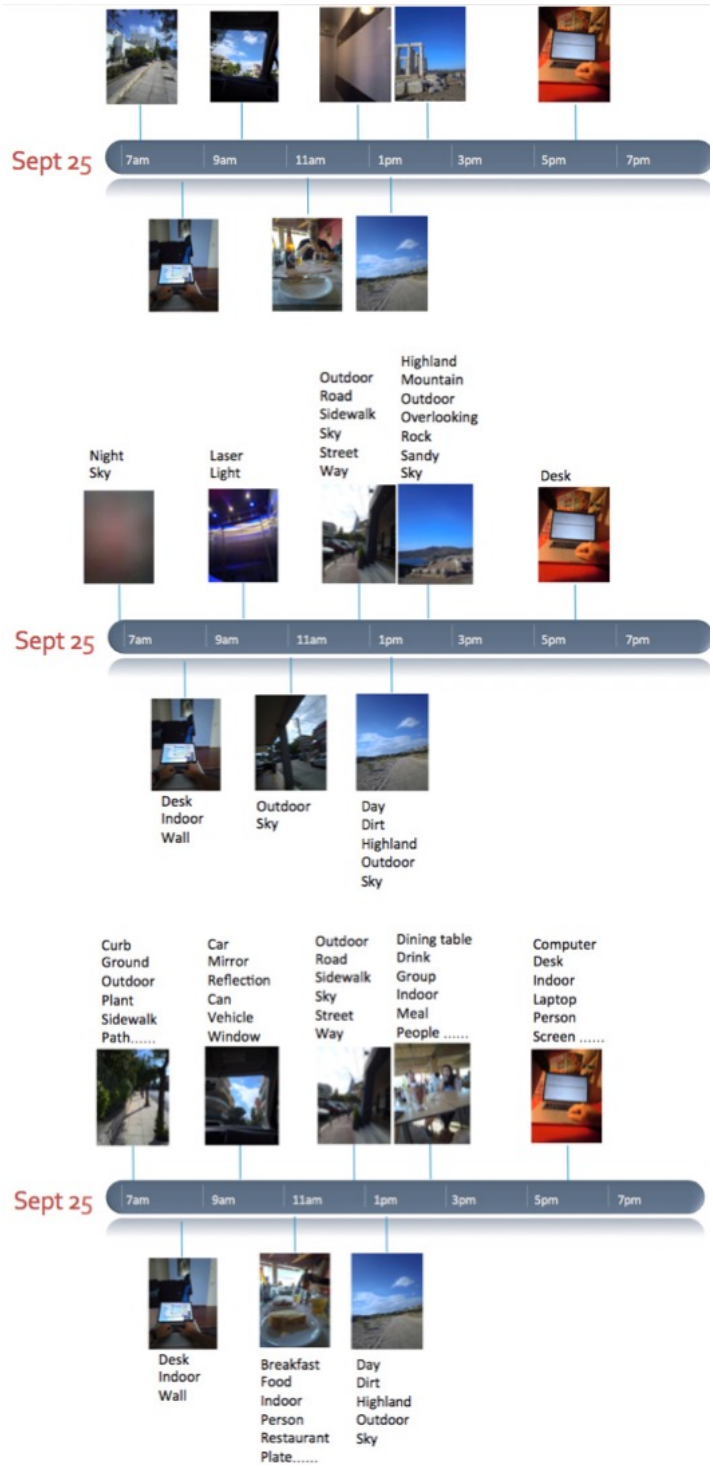

Fig. 4: (a) Aesthetics only. (b) Semantics. (c) Semantics plus some concepts.

Fig. 5: (a) Aesthetics with semantics. (b) Aesthetics, Semantics and some concepts.