Interactive Learning-Based Retrieval Technique for Visual Lifelogging

Ergina Kavallieratou, Carlos R. del-Blanco, Carlos Cuevas, and Narciso García

Abstract. Currently, there is a plethora of video wearable devices that can easily collect data from daily user life. This fact has promoted the development of lifelogging applications for security, healthcare, and leisure. However, the retrieval of not-pre-defined events is still a challenge due to the impossibility of having a potentially unlimited number of fully annotated databases covering all possible events. This work proposes an interactive and weakly supervised learning approach that is able of retrieving any kinds of events using general and weakly annotated databases. The proposed system has been evaluated with the database provided by the Lifelog Moment Retrieval (LMRT) challenge of ImageCLEF (Lifelog2018), where it reached the first position in the final ranking.

Keywords: Lifelogging · Deep learning · Interactive · Weakly annotated · Event detection

1 Introduction

Wearable video cameras are omnipresent in the current consumer market, which has been steadily growing in recent years. Market studies predict that the number of sales will reach 30 million units by 2020 [1]. One of the main keys has been the affordability of wearable cameras, which allow users to continuously record large amounts of unconstrained video data from a first-person point of view, without compromising their mobility or the use of the hands. These facts have promoted video lifelogging, where a user continuously records his everyday experiences by wearing a camera over a long period of time. The acquired images can be exploited to get very useful information about how people live, opening new opportunities for a wide range of applications, such as security, healthcare, and leisure [2–5].

However, the huge amount of image and video data may cause the user never revisits most of those recorded visual memories. Even more, the few relevant events for the user can be extremely difficult to find among long uninteresting segments and repetitive images. Therefore, wearable video devices, and more specifically visual lifelogging, require the development of advanced analysis techniques to identify and locate those meaningful and interesting events and memories. And thus, allowing a fast and efficient data browsing and retrieval.

As a result, the number of research articles and events have increased in the last years to find solutions to the previous demands, such as LifeLog [6], MyLifeBits [7, 8], NTCIR Lifelog Task [9], and the several editions of ImageCLEF Lifelog [10–12]. Most of the existing techniques for video segmentation, summarization, retrieval, and browsing are oriented to Third Person View (TPV) recordings, instead of First Person View (FPV) ones. The analysis of TPV recordings benefits from the existence of constraints imposed by the application domain (sports, news, movies, TV dramas, music videos, etc.) [13, 14]. For example, they rely on flash lights or "score" cuts, background music, shot duration and silences, text captions in broadcast news and shows, etc. [15]. However, these cues are absent in FPV video [16, 17]. Even more, the lack of an intentional structure in the FPV recordings is a source of additional challenges: long streams of data with subtle temporal and spatial boundaries, low quality of the recordings, unknown and diverse context, large number of non-informative images (such as walls or the sky), etc. Consequently, applying TPV analysis techniques to FPV videos is far for providing satisfactory results, even they can perform worse than uniform sampling in some cases [18].

There are also additional challenges in visual lifelogging that restrict the type of visual analysis techniques that can be applied. Some of them makes unreliable the use of computer vision techniques based on temporal coherence and motion estimation, such as the free motion of the camera, the abrupt changes in lighting conditions, and the repetitive image content. Other problems affect the recognition capability of objects in the video, such as occluded objects, blurring, and light saturation [19]. Moreover, the huge volume of data generated by these wearable cameras, along with the current increasing rate of available devices, requires of efficient methods to extract and locate relevant content [20].

Several articles have been proposed in the literature to face the previous challenges for event retrieval and content search in visual lifelogging. Aghazadeh et al. [21] retrieve relevant scenes and actions using a previously acquired egocentric dataset. For this purpose, a query sequence is aligned with sequences in the dataset through dynamic time warping. In [22], visual lifelog data is split into segments, extracting time data, low visual features, and audio features per segment. Then, the user provides a time reference and a query image to extract representative clips per segment by using a clustering approach. Finally, the user can provide additional query images to refine the search and improve the results. Other research proposes to use a more semantic representation, instead of low visual features. In this line, Wang and Smeaton 23 proposed to reason on semantic networks using a density-based approach to extract the most appropriate concepts for event representation. In 2425, a dataset of egocentric images is represented by a graph, adding connections between nodes when the underlying images have a similar Bag-of-Words representation. Finally, a local graph-clustering strategy is applied to retrieve the desired information. Instead of providing a query image, Radeva et al. [26] proposed to measure the similarity between daily visual data, combining dynamic time warping and the Swain's distance. Penna et al. [27] proposed a generative model to capture the feature distribution in video data using deep features. Then, Markov [28] walks are applied over a model that captures the spatial interdependence of the image features. This allows to classify scenes with few labeled training examples.

In this paper, a strategy to retrieve events not previously pre-defined from a huge lifelogging dataset without ground truth is presented. The proposal is based on an interactive and weakly supervised learning approach, where a few query images are required (between 6 and 12). By using semantic image representations based on deep features, a set of related images to the query ones is obtained from the lifelogging database. The user, then, interactively selects the images closest to his original queries. Automatically, the learning-based engine is re-trained, making new predictions that provide the final retrieval results. This procedure is user-friendly, avoiding the requirement of experts to prepare the retrieval system for obtaining new events. This approach has been evaluated in the LMRT challenge of imageCLEF 2018, reaching the first place (out of a total of 29 strategies proposed by 6 different teams).

The rest of the paper is organized as follows. Section 2 describes in brief the LMRT challenge. In Sect. 3 the developed challenge winning strategies are introduced. Sections 4, 5, and 6 describes the database preprocessing, the main methodology after the proposed strategies, and the postprocessing, respectively. Section 7 presents the experimental results. Finally, conclusions are drawn in Sect. 8.

Table 1. Topics considered in the LMRT challenge.

Topic ID	Topic title
LST001	Preparing salad
LST002	VR experiments
LST003	My presentations
LST004	Interviewed by a TV presenter
LST005	Dinner at home
LST006	Assembling furniture
LST007	Taking a coach/bus in foreign countries
LST008	Costa coffee with friends
LST009	Using mobile phone or tablets in a vehicle
LST010	Graveyard

2 Lifelog Moment Retrieval (LMRT) Challenge

The aim of the LMRT challenge is to retrieve specific moments in a lifelogger's life for the 10 topics shown in Table 1. Such moments are defined as semantic events or activities that happened throughout the day.

The provided lifelogging dataset is composed by 50 days of data from a lifelogger. The data can be divided into images, visual concepts, and semantic content. The image data consists of 1500–2500 images per day, acquired from a wearable camera. Visual concepts are automatically extracted with varying rates of accuracy. Regarding semantic content, it is composed by locations, activities, and biometrics information (heart rate, galvanic skin response, calorie burn, steps, etc.), obtained with different sensors and devices. Finally, it must be noted that the dataset does not include specific ground truth related to the specific topics.

3 Proposed Strategies

Three different strategies have been developed for addressing the LMRT challenge. All of them have in common the adoption of a Deep Neural Network-based classification approach that uses an interactive transfer learning method. On the other hand, they differ in the number of simultaneous considered classes (i.e. topics).

The first strategy, called two-class strategy, considers every topic independently, and it requires a trained deep neural network (DNN) per each topic with two outputs: Correct/Wrong. Therefore, since each DNN considers only one topic, each of them will lead to a binary output that represents the topic event or its absence.

The second strategy, called ten-class strategy, considers all the topics simultaneously. Consequently, in this case, only one trained DNN with ten outputs is considered (one output per topic).

Finally, the third strategy, called eleven-class strategy, is an evolution of the previous one with an additional output to consider events that do not belong to any of the 10 topics.

The details concerning each of the three above described strategies are provided throughout the following sections. In addition, the offline pre-processing and post-processing stages that are applied for all the strategies are also detailed.

Table 2. Corresponding images per topic, as they have been described in Table 1.

Topic ID	Category	#images
LST001	Location	27,880
LST002	Activity	66,506
LST003	Activity	66,506
LST004	Location	27,880
LST005	Location	8,986
LST006	Activity	66,506
LST007	Activity	8,800
LST008	Location	601
LST009	Activity	10,754
LST010	Location	26,393

4 Off-Line Preprocessing

According to the types of metadata associated to the images in the dataset, they are first classified into two classes (Activity and Location), and then in several subcategories. For the case of Activity category, the subcategories are: transport, airplane, walking, and no-activity (all the images with no activity information). Regarding the Location category, 96 subcategories are considered, 95 are associated to specific geographical locations, and the last one to images without geographical position information.

This preprocessing will help the user to select images related to specific moments. Moreover, since the process is offline, there is no impact in the computational cost of the proposed strategies.

Table 2 shows how the categories have been assigned to each of the above challenge topics. In addition, the amount of frames conforming each topic is shown.

5 Methodology

The proposed retrieval strategies are based on a six-stage methodology, in which the user must adjusts a pre-trained DNN in an interactive, easy, and fast way to recover the required information of the events. The six stages in the proposed methodology, which are illustrated in Fig. 1, are described below.

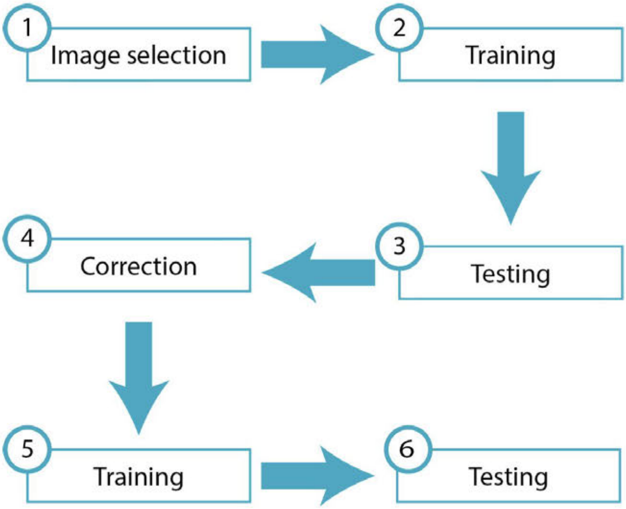

Fig. 1. Block diagram of the proposed classification strategies.

1-Image Selection: First, helped by the classifications performed in the pre-processing stage, the user manually selects sets of images corresponding to each of the topics to be retrieved (true samples). Additionally, for each topic, a second set of images not related to it is also manually created (false samples). Table 3 summarizes the number of positive and negative samples manually selected to address each of the topics. It can be observed that different amounts of true and false samples have been chosen in each topic. To prove that this manual step is not critical, and it can be done easily and quickly, for each topic, the true images have been chosen from a unique event (i.e. same day and same place). In the case of false samples, they have been just selected from the true images corresponding to other topics. Consequently, as it is shown in the table, some topics contain large amounts of true/false samples (e.g. LST003), whereas other ones include much less samples (e.g. LST008). However, it will be proven that the obtained results are successful independently of this large variety of sample set sizes. Consequently, their manual selection can be performed very easily.

2-Training: Once Positive and Negative sets of samples have been manually chosen, the pre-trained Convolution Neural Networks (CNNs) AlexNet [29] and GoogleNet [30] are retrained using such sets.

3-Testing. For each topic, both CNNs are used to automatically classify the images selected in the previous stage as belonging to a topic or not. Depending on the applied strategy (two-class, ten-class, or eleven-class), this stage is performed considering different number of classes. More details are provided later.

Table 3. Number of initial positive and negative samples per topic.

Topic ID	Positive	Negative
LST001	12	400
LST002	22	431
LST003	26	1201
LST004	10	431
LST005	10	2044
LST006	24	26
LST007	10	26
LST008	9	78
LST009	10	102
LST010	8	691

4-Correction: The results obtained for each topic are supervised and, if necessary, manually corrected. In most topics such results were very successful. So, few minutes were necessary to correct the classification results. However, in some topics (e.g. 006), in which the image selection step was problematic, and the final set of true samples was small, the number of misclassified images was significantly higher. Therefore, much more than a couple of minutes would be necessary to correct such misclassifications. However, to prove that the proposed strategy does not require so hard manual interactions, a maximum time of five minutes has been established to perform this correction step in each category. Obtained results have proven that even if all the results initially obtained have not been corrected, the subsequent stages will be able to correctly reclassify the images.

5-Training: Again, once the results previously obtained have been re-classified, both CNNs are retrained using the new sets of true and false sets of images for each topic.

6-Testing: Each CNN is finally used to classify the full set of original images (i.e. all the 80,439 images, without having made any previous classification among them).

In the case of the two-class strategy, the six described stages are applied independently for each topic. That is, for each topic, two classes are considered: images belonging to the topic (true samples) and images not belonging to it (false samples). Therefore, each CNN is used individually for each topic. The initial true samples are set as those manually classified (stage 1) as positive for the corresponding topic, whereas the initial false positives are those belonging to the remaining ones.

Regarding the ten-class strategy, in contrast to the previous one, the CNNs are used to classify the images simultaneously among ten classes, each of them corresponding to each of the topics. Therefore, in this case, in the initial manual classification the samples corresponding to each class are those set as positive (in the corresponding topic) in the pre-processing stage.

Finally, the eleven-class strategy also considers the ten topics simultaneously. However, in contrast to the ten-class strategy, an eleventh class is considered, which includes images not belonging to any of the topics. The samples for this class results from the union of all the groups of negative samples that have been obtained in the pre-processing stage, but discarding those samples belonging to other events.

6 Postprocessing

Once the final classifications have been performed, according to the LMRT rules, a set of 50 images representing each of the topics must be provided. Therefore, a final post-processing stage is necessary to select such 50 images from the set of images classified as belonging to each of the topics.

The used CNNs not only provide a final classification but also a confidence score for each analyzed image. Consequently, to select the 50 most representative images of each class, all the images have been ranked according to such score and those with the highest 50 values have been finally selected.

7 Experimental Results

The proposed system has been evaluated with the databases provided by the Lifelog Moment Retrieval (LMRT) challenge of ImageCLEF (Lifelog2018), where it reached the first position in the final ranking. At the competition, the organizers proposed the classic metrics for retrieval, specifically:

- Cluster Recall at X (CR@X) - a metric that assesses how many different clusters from the ground truth are represented among the top X results;

- Precision at X (P@X) - measures the number of relevant photos among the top X results;
- F1-measure at X (F1@X) - the harmonic mean of the previous two.

Table 4. Indicative results of F1@10 for the proposed strategies. subm#0 has not been submitted to the challenge.

Submission ID	Strategy	CNN	F1@10
subm#1	Two-class	AlexNet	0.504
subm#2	**Two-class**	**GoogleNet**	**0.545**
subm#3	Two-class	Average	0.477
subm#4	Ten-class	AlexNet	0.536
subm#5	Ten-class	GoogleNet	0.477
subm#6	Eleven-class	AlexNet	0.480
subm#0	Eleven-class	GoogleNet	0.542

Table 5. All the results of the competition. DCU was given as reference by the organizers.

Group Name	F1@10	Rank F1@10
AILabGTi	**0.545**	**1**
HCMUS	0.479	2
Regim_Lab	0.424	3
NLP-lab	0.395	4
CAMPUS-UPB	0.216	5
DCU*	0.131	0

All the presented results have been performed using Matlab along with a computer provided with multi-CPU system at 2.80 GHz and a GPU. The mentioned results were also provided by the organizers, off competition.

Official ranking metrics this year is considered the F1-measure@10, which gives equal importance to diversity (via CR@10) and relevance (via P@10). In Table 4, indicative results of F1@10 are given for all the submissions (subm#1-6), plus the not-submitted trial of the third strategy (subm#0). Thus, formally the best strategy proved to be the two-class strategy with the GoogleNet pretrained network. In Table 5, the formal best result of the subtask for every team is presented. Please notice that the runs of DCU* are not ranked since they are the organizing team.

In Table 6, F1@X for various cut off points are considered, with X = 5, 10, 20, 30, 40, 50, for all the submissions.

In Fig. 2, the resulted F1@X are presented in chart per submission. As it is apparent, in most submission the result is not significantly changing by checking more data. As only exception, at subm#1: two-class Alexnet, considering more data improves significantly the result, and it reaches to be much better than the other submissions.

In Fig. 3, the F1@10 is presented per topic. Here, more conclusions can be extracted:

- For the topics LST003, LST004 and LST006, the results do not change by the different techniques. More obvious is the case of LST006, where it is always 0, since the initially selected images were wrong examples.
- It seems that in subm4: ten-classes Alexnet, most of the queries present a peak, except of LST002 that presents low.
- It is interesting that LST004: Interviewed by a TV presenter, gives almost perfect results. Since there were many images in different places, and just few were selected at first, could it be that the presence of the camera is enough to distinguish the moments?

Table 6. Results for all the trials of F1@X for X = 5, 10, 20, 30, 40, 50. subm#0 has not been submitted to the challenge.

Trial	F1@5	F1@10	F1@20	F1@30	F1@40	F1@50
sub#1	0.395	0.504	0.571	0.604	0.606	0.594
sub#2	**0.520**	**0.545**	**0.562**	**0.547**	**0.523**	**0.522**
sub#3	0.452	0.477	0.445	0.438	0.465	0.473
sub#4	0.543	0.536	0.543	0.552	0.562	0.556
sub#5	0.452	0.477	0.459	0.438	0.465	0.473
sub#6	0.480	0.480	0.495	0.521	0.528	0.549
sub#0	0.507	0.542	0.525	0.534	0.508	0.532

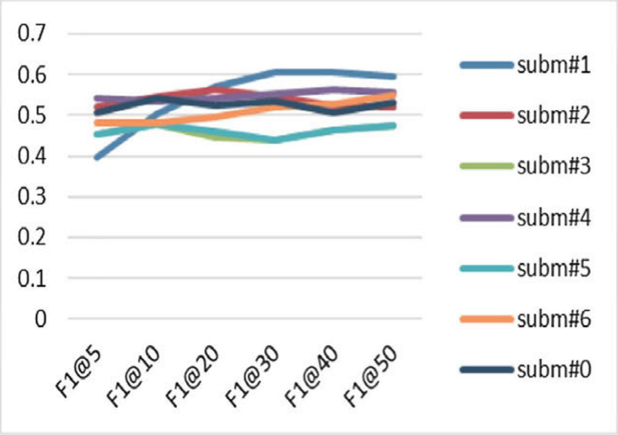

Fig. 2. F1@X per submission.

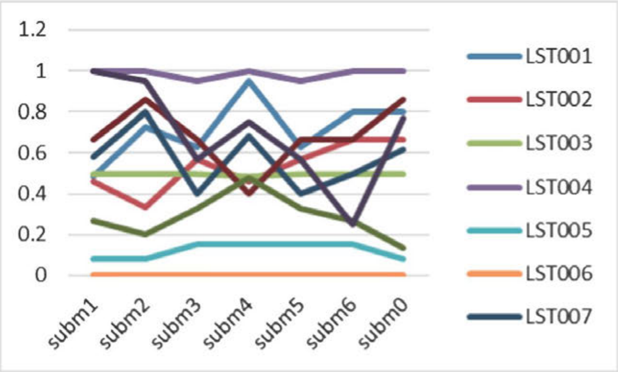

Fig. 3. F1@10 per topic.

8 Conclusions

This paper describes an interactive and weakly supervised learning system that is able of retrieving any kinds of events using general and weakly annotated databases. It was evaluated in the framework of the Lifelog Moment Retrieval (LMRT) challenge of ImageCLEF Lifelog2018, that it came first in the final ranking [31]. The competition was quite challenging as it required to handle a huge number of images for retrieving moments for ten specific topics. We proposed 3 different strategies to respond to the topics, all using deep learning-based algorithms and specifically AlexNet and GoogleNet.

References

1. Wearable Cameras: Global Market Analysis and Forecasts, Tractica, Boulder, CO, USA, (2015)
2. Jalal, A., Uddin, M.Z., Kim, T.S.: Depth video-based human activity recognition system using translation and scaling invariant features for life logging at smart home. IEEE Trans. Consum. Electron. **58**(3), 863–871 (2012)
3. Doherty, A.R., et al.: Experiences of aiding autobio- graphical memory using the sensecam. Hum.-Comput. Interact. **27**(1–2), 151–174 (2012)
4. Hodges, S., et al.: SenseCam: a retrospective memory aid. In: Dourish, P., Friday, A. (eds.) UbiComp 2006. LNCS, vol. 4206, pp. 177–193. Springer, Heidelberg (2006). https://doi.org/10.1007/11853565_11
5. Lee, M.L., Dey, A.K.: Lifelogging memory appliance for people with episodic memory impairment. In: Proceedings of the 10th International Conference on Ubiquitous Computing, pp. 44–53. ACM (2008)
6. Magazine, G.: LifeLog: DARPA looking to record lives of interested parties (2013). https://www.geek.com/news/lifelog-darpa-looking-torecord-lives-of-interested-parties-552879/. Accessed 28 May 2018

7. Gemmell, J., Bell, G., Lueder, R., Drucker, S., Wong, C.: MyLifeBits: fulfilling the Memex vision. In: Proceedings of the Tenth ACM International Conference on Multimedia, pp. 235–238. ACM (2002)
8. Gemmell, J., Bell, G., Lueder, R.: MyLifeBits: a personal database for everything. Commun. ACM **49**(1), 88–95 (2006)
9. Gurrin, C., Joho, H., Hopfgartner, F., Zhou, L., Albatal, R.: Overview of NTCIR-12 lifelog task. In: Proceedings of the 12th NTCIR Conference on Evaluation of Information Access Technologies, Tokyo, Japan (2012)
10. Dang-Nguyen, D.T., Piras, L., Riegler, M., Boato, G., Zhou, L., Gurrin, C.: Overview of ImageCLEFlifelog 2017: lifelog retrieval and summarization. In: CLEF2017 Working Notes, Dublin, Ireland, vol. 1866 (2017)
11. Dang-Nguyen, D.T., Piras, L., Riegler, M., Zhou, L., Lux, M., Gurrin, C.: Overview of ImageCLEFlifelog 2018: daily living understanding and lifelog moment retrieval. In: CLEF2018 Working Notes. CEUR Workshop Proceedings (2018)
12. Ionescu, B., et al.: Overview of ImageCLEF 2018: challenges, datasets and evaluation. In: Bellot, P., et al. (eds.) CLEF 2018. LNCS, vol. 11018, pp. 309–334. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-98932-7_28
13. Gygli, M., Grabner, H., Van Gool, L.: Video summarization by learning submodular mixtures of objectives. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3090–3098 (2015)
14. Lin, Y.-L., Morariu, V., Hsu, W.: Summarizing while recording: context-based highlight detection for egocentric videos. In: Proceedings of the IEEE International Conference on Computer Vision Workshops, pp. 51–59 (2015)
15. Money, A.G., Agius, H.: Video summarisation: a conceptual frame- work and survey of the state of the art. J. Vis. Commun. Image Represent. **19**(2), 121–143 (2008)
16. Bolanos, M., Dimiccoli, M., Radeva, P.: Towards storytelling from visual lifelogging: an overview, arXiv preprint arXiv:1507.06120 (2015)
17. Betancourt, A., Morerio, P., Regazzoni, C.S., Rauterberg, M.: The evolution of first person vision methods: a survey. IEEE Trans. Circ. Syst. Video Technol. **25**(5), 744–760 (2015)
18. Lee, Y.J., Grauman, K.: Predicting important objects for egocentric summarization. Int. J. Comput. Vis. **114**, 38–55 (2015)
19. Tan, C., Goh, H., Chandrasekhar, V., Li, L., Lim, J.H.: Understanding the nature of first-person videos: characterization and classification using low-level features. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 549–556. IEEE (2014)
20. Bolanos, M., Dimiccoli, M., Radeva, P.: Toward storytelling from visual lifelogging: an overview. IEEE Trans. Hum.-Mach. Syst. **47**(1), 77–90 (2017)
21. Aghazadeh, O., Sullivan, J., Carlsson, S.: Novelty detection from an ego-centric perspective. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 3297–3304 (2011)
22. Wang, Z., Hoffman, M.D., Cook, P.R., Li, K.: 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)
23. Wang, P., Smeaton, A.F.: Semantics-based selection of everyday concepts in visual lifelogging. Int. J. Multimedia Inf. Retrieval **1**(2), 87–101 (2012)
24. Min, W., Li, X., Tan, C., Mandal, B., Li, L., Lim, J.H.: Efficient retrieval from large-scale egocentric visual data using a sparse graph representation. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 541–548 (2014)
25. Chandrasekhar, V., Tan, C., Min, W., Liyuan, L., Xiaoli, L., Hwee, L.J.: Incremental graph clustering for efficient retrieval from streaming egocentric video data. In: IEEE International Conference on Pattern Recognition, pp. 2631–2636 (2014)

26. Radeva, P., Aksasse, B., Ouanan, M.: Using content-based image retrieval to automatically assess day similarity in visual lifelogs. In: 2017 Intelligent Systems and Computer Vision (ISCV). IEEE (2017)
27. Penna, A., Mohammadi, S., Jojic, N., Murino, V.: Summarization and classification of wearable camera streams by learning the distributions over deep features of out-of-sample image sequences. In: IEEE International Conference on Computer Vision (ICCV), Venice, pp. 4336–4344 (2017)
28. Rabiner, L.R.: A tutorial on hidden Markov models and selected applications in speech recognition. Proc. IEEE **77**(2), 257–286 (1989)
29. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, pp. 1097–1105 (2012)
30. Szegedy, C., et al.: Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–9 (2015)
31. http://imageclef.org/2018/lifelog. Accessed 25 Aug 2018