How to Browse through my Large Video Data? Face Recognition & Prioritizing for Lifelog Video

Katrin Wolf[1], Yomna Abdelrahman[2], Mathias Landwehr[2], Geoff Ward[3] and Albrecht Schmidt[2]
[1]Hamburg University of Applied Science, [2]University of Stuttgart, [3] University of Essex
[1]katrin.wolf@acm.org, [2]firstname.lastname@vis.uni-stuttgart.de, [3]gdward@essex.ac.uk

ABSTRACT

Due to the rise of lifelog cameras, we have personal video data that is too large to be watched. Video indexing has the potential to provide meta-information for faster video search. This work aims to support lifelog video indexing through automated face priority rating. In a user study, we identified parameters that allow for rating the importance of persons in a video. We implemented these findings to automatically predict the person's importance in video. We show that our algorithm predicts similar person priority ratings like the participants had given. Hence, we contribute to video-based lifelogging through indicating, implementing, and testing face indexing rules that predict how important a person in a video is perceived. Our findings can help to build video players that support users navigating through their large video data and reviewing sequences that recall important moments of life.

ACM Classification Keywords

H.5.2 User Interfaces: Graphical user interfaces.

Author Keywords

Video navigation; lifeloging video; face detection.

BACKGROUND

Due to the rise of wearable cameras, lifelogging, the process of automatically recording aspects of one's life in digital form [6], is producing a massively growing amount of both, image and video data. Navigating through such large video, e.g., for re-viewing a specific life scene, is very time consuming and is not yet sufficiently supported by software. This work aims to support lifelog video navigation through automatically providing useful meta-data for video indexing.

We argue that we should aim to design lifelog video indexing algorithms as similar as possible to strategies humans use to access their autobiographical memory. It is widely accepted that autobiographical memories of past situations can be accessed by a wide variety of cues, such as what happened, who was there, when it took place, and where it occurred (e.g., [2], [4], [14]). It is generally observed that providing "what" happened is the most effective cue to the past event, with "who" and "where" also highly informative relative to "when" ([4], [14]), and that multiple cues are more effective than single cues ([14]), e.g., recalling what happened is greatly improved by providing both "who" and "where" compared with providing just one or the other. Inspired by the autobiographical memory process, we propose to add information about "who", "where" and "when" as meta-data to video sequences. While many current cameras already save the "when" as time stamp and the "where" as GPS data; the information about "who" was captured is missing in video data so far as well as how important the "who" may be perceived in the video. For personal photo sorting and finding, computer aided support, like iPhoto or the face recognition of Facebook, is a beneficial automated software approach to better and faster navigate through photo collections. While these algorithms work well for consciously taken photos, lifelog images, due to the nature of wearable cameras, usually contain much unimportant information and suffer from bad light conditions and motion blur [15]. This paper contributes to lifelog video navigation through proposing an automated aid that adds information about the person shown in the video and about his/her's importance as such information would allow for faster navigating through lifelogging video.

Although automated solutions have been developed for lifelog photo activity [6] and event [9] classification, to date no automated solutions exist for indexing lifelogging video. However, manual approaches have been proposed for video indexing, but such solutions will not be efficient enough to support the amount of lifelogging video data that we will soon have. For example, Christel [5] added manually semantic text labels about the location shown in video, which is a very time consuming approach. Existing automated solutions are adding meta-information to photos about time and place. Moreover, face detection and recognition has been used for photo indexing. For example, in iPhoto or Facebook, faces are automatically highlighted, and then users can manually add the name to a face or they can link a face to a certain Facebook profile. For automated video indexing, Ma and Zhang added location information to video through GPS data of the camera [10]. Al-Hajri et al. proposed a video-watch-history approach for scene importance indexing, considering more often watched videos are more important (similar to the recommender system of youtube) [1]. Gao et al. [7] proposed an importance ranking for people in TV series and movies using a face recognition algorithm. Here the aim was to find the main cast, which was achieved by the assumptions that main characters have more screen time than others. In order to index general lifelog data beyond video, Gemmel et al. [8] introduced a database application where as many as possible actions of the user is saved. The aim here is to give maxi-

MUM'16, December 12–December 15, 2016, Rovaniemi, Finland.
Copyright © 2016 ACM ISBN/978-1-4503-4860-7/16/12 $15.00.
http://dx.doi.org/10.1145/3012709.3012713

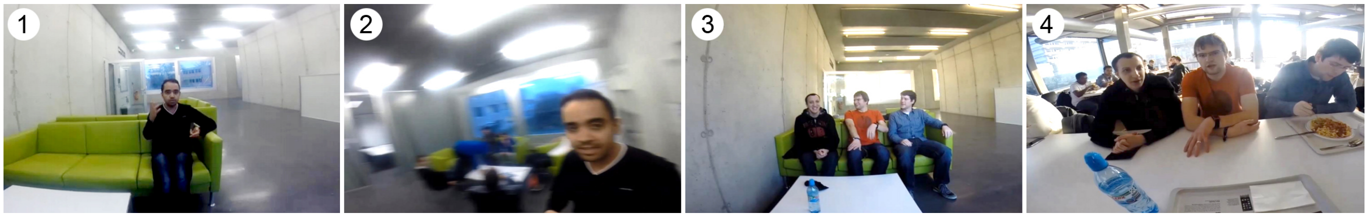

Figure 1. Video scenarios: dialogue (1), moving dialogue (2), meeting (3), and eating (4).

mum amount of different context information to help the user remember past actions. This application saved every context information available, e.g., mouse clicks on the computer or the weather when emails were sent and received.

Solutions to add meta-information to video about the ”when” and the ”where” already exist using time stamp and GPS. As lifelogging generates massively large data, we aim to extend existing video indexing approaches by creating a fully automated process for adding meta-information about the ”who” to lifelog video. We first conducted a user study to examine what parameters of image content would be helpful in determining which faces are worth highlighting. We then used the importance rating parameters to implement a face recognition algorithm, and tested the extent to which such an algorithm could automatically predict who is important in lifelogging video.

EXPERIMENT

We conducted a user study to explore the ”who” in lifelog video. Lifelog video captures all kinds of everyday situations, and, (in contrast to movies), due to the nature of constantly and passively lifelogging, faces randomly appear in the video. Thus, many persons shown in lifelog video have no importance for the owner of the lifelogging device. In this experiment, we particularly aimed to understand which parameters cause importance of faces in such videos using 4 clips that represent different scenario types in lifelog video, being in a dialogue at a table, sitting with two people at a table, walking while having a dialogue, and a lunch situation. Of course, these four scenarios barely represent lifelog video. However, we believe that they represent typical everyday situations that are likely to be often recorded in lifelog video. We are aware that using somebody else’s lifelog video is artificial. However, using the same video material for all participants has the advantage to guarantee equal conditions for all participants. Moreover, personal videos contain emotional meta-information that would influence our results as stated by Wagenaar: ”Pleasant events were better recalled than unpleasant events” [14]. To avoid that, we use video participants have no emotional connection with.

Task, procedure & measures

For identifying clear indicators that let a face in a video appear to be important and for exploring the underlying parameters of the importance of the face, we produced mentioned 4 lifelog simulating videos that we showed to participants. The viewing order of the videos was arranged using Latin square. Before watching the videos, the participants were asked to imagine that the videos were lifelog data that were recorded to support recalling their past. While watching the video, participants were asked to select faces they thought would be relevant or desirable to recall. The selection was realized through pausing the video and cropping the face with a dedicated tool of our apparatus. For each selected face, we asked participants to rate the importance that the person may have for somebody who wants to recall his/her life using a 7-item Likert scale. To better understand why participants found a person to be important, we also asked participants through open questions to name reasons why they had selected the specific face.

Apparatus

Our videos lasted about 1.5 minutes each and had a frame rate of 30fps. The audio was muted to not distract the visual attention of the participants and also because of privacy issues. The videos (shown in Figure 1) had the following content:

(1) A dialogue between 2 sitting people, one wearing the camera and one sitting opposite.
(2) A dialogue between 2 walking people, one wearing the camera and one walking beside him. Additionally some people were shown in the background.
(3) A group meeting with 4 persons, 3 sitting opposite the person wearing the camera. All 4 were in conversation.
(4) A lunch in a public cafeteria where 3 people were sitting opposite to the person wearing the camera, 2 were in conversation and one was not. Many persons in both, foreground and background.

We implemented a web application for playing back the video and for letting the participants pause the video, select a frame, crop the faces, rate their importance, and name reasons for the selection. The application contained a media player to play/pause the video, a frame selection button, and a photo gallery where the cropped faces were shown. A window popped up when a frame was selected. It showed the frame and allowed the participants to crop the face. Additionally, a Likert scale with radio buttons was presented to rate the face’s importance, and the participants were asked to enter in a text field the reason(s) for the face selection.

Design

Our study had a within subjects design with 16 participants (9 males, 7 females), aged between 23 and 76 (mean=42, SD=19.6). The independent variable was the video content. Each of the 4 videos showed a different scenario with varying persons in different social interactions. The dependent

Index	Category name	Explanation	# usages	# participants
1	Screen time	If a person is all the time on the video	68	10
2	Appearance frequency	When a person appears frequently	15	15
3	Conversation partner	When the person is talking to the camera	104	16
4	Hand gestures	Using hand gestures when talking	30	11
5	Single person	When a person is the only one on the video	10	10
6	Known person	Indicates, if the person is known by person that wears the camera	14	6
7	Unknown person	Indicates if the person is known by the rating person	7	3
8	Center of screen	If a person is in the center of the video frames	5	3
9	Attention-grabbing appearance	If the person wears something coloful	4	2
10	Activity	When a person is doing something	8	8
11	Eye contact	When a person is looking at the camera	13	4
12	Facial expression	Includes laughing, grinning, and other emotions	2	1
13	Close to the camera-wearing person	When a face has much frame size	9	9

Table 1. Categories of reasons that were used for selecting faces.

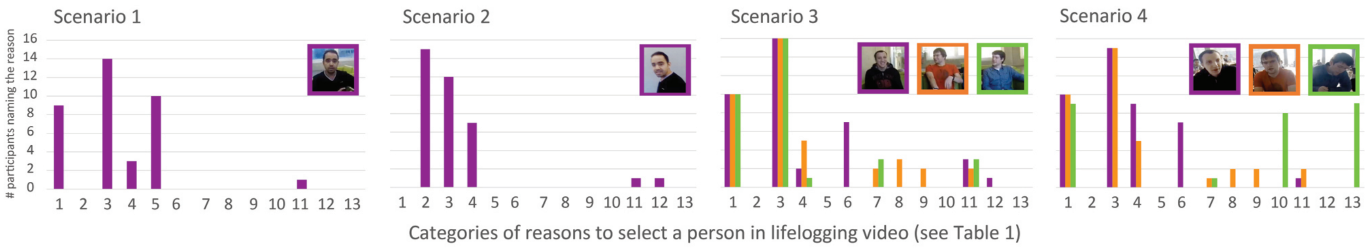

Figure 2. Category ratings per reason given for select a person from the videos in each of the four scenarios.

variables were selected faces, importance rankings for each selected face, and reasons for the face selection.

RESULTS

In the 4 videos, the faces of the 8 persons mentioned in the description of the scenarios were selected, while faces in the background were never chosen, see Figure 3. Even the face of a person not interacting with the person that wore the camera (scenario 4, face 3) was only chosen with a low importance rating. Different reasons for the selection were provided, and sometimes more than one reason per selection was given. Using a bottom-up analysis and open coding, we grouped the selection reasons according to their semantic closeness and their appearance frequency, see Table 1. We also reported how many participants mentioned a particular reason.

The reason categories were further used to show the reasons per face and scenario that led to the selection, as shown in Figure 2. Hence, we analyzed why a certain person in a particular scenario was selected. An important reason for the selection of a person is his/her screen time (Table 1: category 1, 2). Moreover, in all scenarios, only persons in the foreground were selected to be important (see Figure 3 and Table 1: category 5, 8, 13), while the importance depends to a great extent on the behavior of the person. The person is considered to be very important if being in conversation with the user (Table 1: category 3, 4, 11, 13). This can be inferred by mouth movements and through gesticulation. The importance is constantly decreasing with a reduction in conversation activity. If a none-communicating person is selected, the importance may be very low (see Figure 3: scenario 4, person 3 and Table 1: category 13). The selection then may have other reasons, e.g. he did something (Table 1: category 10).

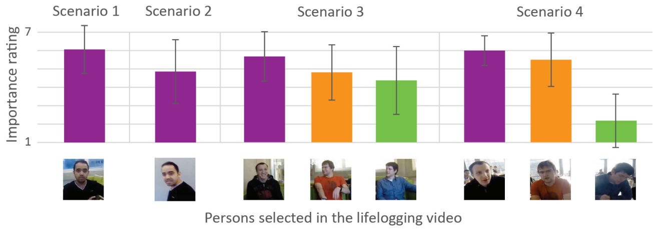

Figure 3. Persons' importance rated by participants (mean, SD).

DISCUSSION

Through showing lifelog simulating videos to participants, we identified reasons why persons in lifelog video would be worth indexing for later situation recall. Greater screen time and higher frequency of appearance has been observed to predict the importance of a person. However, our participants found only persons in the foreground important; but the perceived importance of the foreground faces differed.

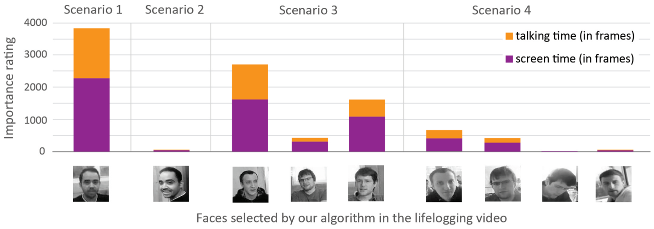

Figure 4. Importance ratings per face detected by our algorithm.

Most important were people rated that were communicating with the person that wore the camera. This was inferred by mouth movements, gesticulation, eye contact, and emotional mimic, e.g. grinning. Persons that are not communicating with the lifelogging person were rated as much less important; but they may still be interesting, e.g., if being active.

Our findings are in line with whose of Gao et al. [7] who proposed an importance ranking for people in TV series and movies based on their screen time. Thus, we can transfer this face importance indicator from TV productions to lifelog video. Moreover, we extend the list of importance indicators by also recommending to consider whether the person is in the foreground (as major indicator) and if the person is in conversation (to distinguish the importance). Conversation activities can be identified by mouth movements, gesticulations, mimic, and eye contact. Other activities may also indicate some importance. Speech and audio analysis would further improve the understanding the importance of people in video; but analyzing people's conversation causes much more privacy issues than considering only the images of video.

IMPLEMENTATION & TESTING

Here, we describe the implementation of an automated face detection and recognition algorithm for lifelog video that provides video indexing information about the importance of the recognized faces considering the findings of our experiment. This serves both as a proof-of-concept and as a test to determine whether an algorithm considering screen time and communication can get similar results than those we got from our user study. Hence, we used our 4 experimental videos.

Like Gemmel et al. [8], we used a database to store the detected faces (due to the algorithm output in gray scale) as well as the meta-information, including screen time (as we confirmed the assumption of Gao et al. [7] that screen time is related to the importance of a face), size (to indicate closeness, which also indicates importance), and conversation time (as we found that for shown faces, the ones that are communicating are more important than the others). The data base allows for storing time and place if the camera has a GPS sensor, which we recommend to also consider for adding meta-information to lifelog video about ”where” and ”when”.

For the face detection, recognition, and importance rating the following steps were performed for each frame: In order to boost the processing, the video size is reduced to 800x600 pixels. To detect and recognize faces we used the OpenCV Viola Jones Algorithm [13], the eigenface recognizer [12], and variations of the Viola Jones Algorithm by using different cascade classier in order to identify faces with a prole view. We systematically tested the recognition accuracy with different face recognizers including the local binary patterns [11], Fisherfaces [3], and eigenface recognizer. Viola Jones Algorithm and the eigenface recognizer showed the best recognition accuracy. The detected faces automatically got a face ID, and their frames time stamps were saved in the data base immediately after the recognition. Due to head movements or lightning noise, a face can disappear for a moment when still being part of the scene. Hence, we defined a threshold for time gaps between two faces appearances of 0.3 sec (10 frames, with 30fps) for that we assume that the face also was present in the frames between. Multiple faces can be detected and recognized in one frame. Our data base contained a number of false positives. Skin detection was used to reduce the number of false positives, and a face size threshold excluded faces shown in the background (see category 13 of Table 1). A final manual selection served for filtering our results.

Through that procedure we detected 9 faces in the 4 video clips. For them, we calculated the importance rating using screen time and communication activity. The absolute screen time a person had was represented through the amount of frames the face was recognized. For each frame that showed a face, we set a Boolean value to 1 if the person was talking and to 0 if not. We detected talking activity using a histogram comparison algorithm for the mouth region for every 5 following frames, which allowed for detecting mouth movements in a frame sequence. The Boolean values of the talking indication are summed up to calculate the total frame number of talking activity. The importance of a face is then calculated as the sum of all talking frames and the screen time itself.

Comparing the results of our user study with the results of our algorithm, we see that the user-defined important persons in our study overlap with 8 out of the 9 detected faces, see Figure 4. Similar to the previous study, the 7 persons that were talking to the person with the camera got highest importance scores. The only exception was the person walking in scenario 2 beside the lifelogging person. As the person mainly focused on the way, the conversation partner does not often occur and thus, got only short screen time and consequently little conversation time as well. The calm person that our participants had selected was also detected by our algorithm and he, similar to the user-rating, got rather low importance values. Our algorithm detected one person that was not selected by participants: a man on the very left side of the video capture and who is also not talking into the camera but to somebody else. Reviewing the importance categories defined in our user study, we could consider here category 8 of Table 1 that recommends to only consider persons in the center. Hence, an improved version of our algorithm will not consider face at the frame border. Finally, we propose to introduce a concept of presence for persons that are there but not captured, like the one walking beside the main person.

CONCLUSION

We found that screen time, face size, and communication activities are used in people's judgment about the importance of persons in lifelog video. We implemented a simple algorithm that predicts the importance of persons in such video using the parameters identified in our study. A test of our algorithm shows that our algorithm successfully identified the same persons as our participants did and rates their importance mostly similarly. Hence, our algorithm approach can help to build video browsers that automatically highlight important scenes, which is crucial for organizing and navigating through large video, such as captured with lifelogging cameras. A larger-scale batch testing could confirm the scalability of our results as well as also considering female protagonists. Future work could benefit from our work by applying our approach for object detection in lifelogging video.

REFERENCES

1. Abir Al-Hajri, Gregor Miller, Matthew Fong, and Sidney S. Fels. 2014. Visualization of Personal History for Video Navigation. In *Proceedings of the 32Nd Annual ACM Conference on Human Factors in Computing Systems (CHI '14)*. ACM, New York, NY, USA, 1187–1196. DOI: http://dx.doi.org/10.1145/2556288.2557106

2. Lawrence W Barsalou. 1988. The content and organization of autobiographical memories. *Remembering reconsidered: Ecological and traditional approaches to the study of memory* (1988), 193–243.

3. Peter N. Belhumeur, João P. Hespanha, and David J. Kriegman. 1997. Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection. *IEEE Trans. Pattern Anal. Mach. Intell.* 19, 7 (July 1997), 711–720. DOI: http://dx.doi.org/10.1109/34.598228

4. Christopher DB Burt. 1992. Retrieval characteristics of autobiographical memories: Event and date information. *Applied Cognitive Psychology* 6, 5 (1992), 389–404.

5. Michael G. Christel. 2008. Supporting Video Library Exploratory Search: When Storyboards Are Not Enough. In *Proceedings of the 2008 International Conference on Content-based Image and Video Retrieval (CIVR '08)*. ACM, New York, NY, USA, 447–456. DOI: http://dx.doi.org/10.1145/1386352.1386410

6. Aiden R Doherty, Niamh Caprani, Ciarán Ó Conaire, Vaiva Kalnikaite, Cathal Gurrin, Alan F Smeaton, and Noel E OConnor. 2011. Passively recognising human activities through lifelogging. *Computers in Human Behavior* 27, 5 (2011), 1948–1958.

7. Yong Gao, Tao Wang, Jianguo Li, YangZhou Du, Wei Hu, Yimin Zhang, and HaiZhou Ai. 2007. Cast Indexing for Videos by NCuts and Page Ranking. In *Proceedings of the 6th ACM International Conference on Image and Video Retrieval (CIVR '07)*. ACM, New York, NY, USA, 441–447. DOI: http://dx.doi.org/10.1145/1282280.1282345

8. Jim Gemmell, Gordon Bell, and Roger Lueder. 2006. MyLifeBits: A Personal Database for Everything. *Commun. ACM* 49, 1 (Jan. 2006), 88–95. DOI: http://dx.doi.org/10.1145/1107458.1107460

9. Hyowon Lee, Alan F Smeaton, Noel E OConnor, Gareth Jones, Michael Blighe, Daragh Byrne, Aiden Doherty, and Cathal Gurrin. 2008. Constructing a SenseCam visual diary as a media process. *Multimedia Systems* 14, 6 (2008), 341–349.

10. He Ma, Roger Zimmermann, and Seon Ho Kim. 2012. HUGVid: Handling, Indexing and Querying of Uncertain Geo-tagged Videos. In *Proceedings of the 20th International Conference on Advances in Geographic Information Systems (SIGSPATIAL '12)*. ACM, New York, NY, USA, 319–328. DOI: http://dx.doi.org/10.1145/2424321.2424362

11. Ahonen Timo, H Abdenour, and P Matti. 2004. Face recognition with local binary patterns. In *Proceedings of the ECCV*. 469–481.

12. Matthew Turk and Alex Pentland. 1991. Eigenfaces for recognition. *Journal of cognitive neuroscience* 3, 1 (1991), 71–86.

13. Paul Viola and Michael J Jones. 2004. Robust real-time face detection. *International journal of computer vision* 57, 2 (2004), 137–154.

14. Willem A Wagenaar. 1986. My memory: A study of autobiographical memory over six years. *Cognitive psychology* 18, 2 (1986), 225–252.

15. Katrin Wolf, Albrecht Schmidt, Agon Bexheti, and Marc Langheinrich. 2014. Lifelogging: You're Wearing a Camera? *IEEE Pervasive Computing* 13, 3 (2014), 8–12.