Impressive Picture Selection from Wearable Camera Toward Pleasurable Recall of Group Activities

Eriko Kinoshita and Kaori Fujinami(✉)

Department of Computer and Information Sciences, Tokyo University of Agricalture and Technology, 2-24-16 Naka-cho, Koganei, Tokyo 184-8588, Japan
asa.kinoko0216@gmail.com, fujinami@cc.tuat.jp

Abstract. Wearable cameras allow us to capture large amount of video or still images in an automatic and implicit manner. However, the only necessary images should be filtered out from the captured data that contains meaningless and/or redundant information. In this paper, we propose a method to identify a set of still images by audio and video data, which is intended to let users feel pleasurable when they watch the images later.

Keywords: Wearable camera · Life logging · Image and audio analysis

1 Introduction

Wearable cameras such as SenseCam [6], GoPro [3] and A1H [8] enables automatic and implicit life-logging. A user would be aroused a particular emotion when he/she reviews the recorded data by recalling what happened at that moment. However, in such passive life-logging, particular moments should be identified from huge amount of data, e.g., video and still images, to reduce cognitive burden of the user, and summarization techniques have been proposed [1]. The main purpose of existing image summarization techniques is to improve the usability of life log browsing, in which the user's satisfaction in recording and reviewing memories is not fully considered.

We design the image summarization system by taking into account the effect of browsing. More specifically, we aim at detecting a moment in which a group of people feel pleasurable when they review the logged data, which we call *post-pleasurable*. Based on a preliminary user survey with 50 people, we identified two types of post-pleasurable moments obtained from an automatic recording device: (1) the same group members of the photo-taker talking with each other and (2) partying during a group activity. By contrast, beautiful and rare scenes were found to be less meaningful for automatic post-pleasurable scene selection because people explicitly take photos by themselves in such cases. In addition to the two moments, we added (3) "having interests in something" to the target moments because something that attracts a user should make him/her recall special emotions later.

M. Antona and C. Stephanidis (Eds.): UAHCI 2017, Part III, LNCS 10279, pp. 446–456, 2017.
DOI: 10.1007/978-3-319-58700-4_36

In this paper, we propose a method to identify scenes (2) and (3) using audio and visual data from first-person view camera. The rest of the paper is organized as follows. Section 2 examines related work to validate the approach of automatic photo taking and to state our approach in video summarization techniques. In Sect. 3, the system design and implementation is presented. A user study to understand the emotional effects by the proposed system in Sect. 4, followed by discussion in Sect. 5. Finally, we conclude the paper in Sect. 6.

2 Related Work

In the field of life-logging, research focusing on first-person viewpoint using body mounted camera is presented, which is called "visual life-logging" [1]. Sellen, et al. conducted experiments on memory recall using SenseCam [6] as a verification on the effectiveness of memory support of life-logging [10], in which photos taken automatically by a wearable camera is suggested that it is easier to recall past memories than photos taken voluntarily by a still camera. So, the usefulness of automatic shooting in our approach seems to be supported.

Summarization, keyframe selection in other words, techniques from video stream have been proposed to specify particular moments, which are used to reduce cognitive burden of the user to find appropriate ones from huge amount of data [1,2,5,7,9]. Image-based keyframe selection employs visual features such as contrast, color variance, sharpness, noise and saliency to identify non-redundant yet meaningful frame [2,7]. We consider that these visual features are basic ones and that they are not so effective in recalling pleasurable moments. Emotional features are effective in identifying more specific moments that relate to emotions of humans, e.g., enjoyment, fear, surprise, anger, etc. StartleCam [5] is a pioneering work in visual life-logging that leverages electrodermal activity (EDA) sensor, a.k.a. galvanic skin response (GSR) sensor, to extract frightening moments from photo stream, in which the sensor is attached on fingers or foot. Ratsamee, et al. proposed a keyframe selection method based on excitement measured from EDA sensor attached around the wrist, in cooperation with visual features from a smartphone's video camera hanging from the neck [9]. Although their study shares the motivation of keyframe selection from an emotional aspect with ours, we utilize only a video camera to reduce the physical load. Furthermore, we aim at selecting pictures for an entire group members even who do not wear recording devices.

3 System Design and Implementation

3.1 Overview

The proposed system assumes that a user attaches a camera on his/her head, e.g., eye glasses and headphones, and that the judgment process is carried out in an offline manner by providing a movie file. Figure 1 illustrates the processing flow; major components are "partying estimator" and "interest estimator" that

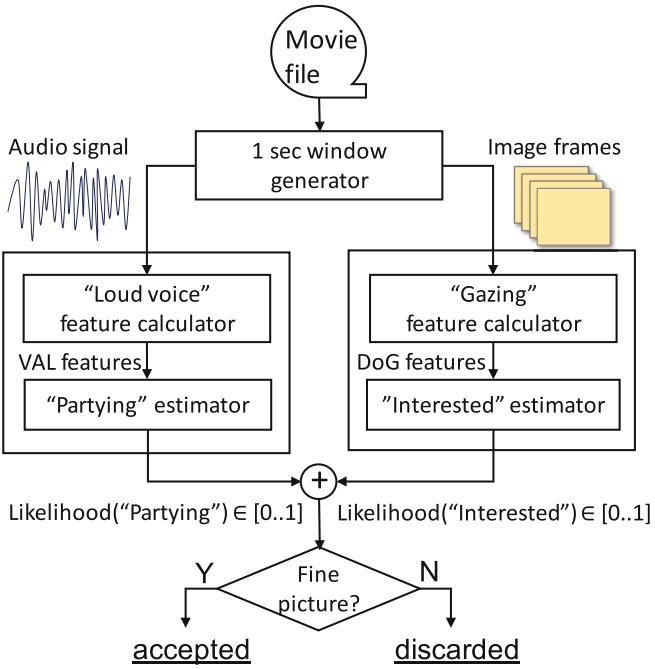

Fig. 1. Processing flow.

corresponds to the moment (2) and (3), respectively. The two estimators consist of binary classifiers that generate the recognition results on whether a particular period of time includes these moments (positive) or not (negative) with accompanying likelihood of *positive* ranging from 0 to 1. The two scores are merged, which is called "fine picture score", and the final decision is made on the score to obtain an output still image.

3.2 Features that Characterize Pleasurable Moments

Voice Audio Level: The "partying" moment is defined as what loud voice turns up, which is estimated every 1 s from audio signals. Voice audio level (VAL) features are represented by absolute audio levels and relative ones.

An absolute voice audio level (A-VAL) is calculated as follows. Firstly, maximum absolute amplitude ($maxAbsAmp$) is found in a sub-window of w ms. Secondly, the moving averages ($aveMaxAbsAmp$s) are calculated against m samples of$maxAbsAmp$s with 1 sample sliding for 1 s. Finally, the sum of ($aveMaxAbsAmp$s) is obtained as an A-VAL. Here, A-VAL$_{w,m}$ represents the value of A-VAL for sub-window of w ms and moving average of m samples. In addition to the absolute value, the difference of consecutive A-VAL$_{w,m}$ is utilized as relative voice audio level (R-VAL$_{w,m}$). By changing the length of sub-window (w), i.e., 100, 500, and 1000 ms, and the number of moving average samples (m), i.e., 7 and 9, a total of 12 features are defined.

Degree of Gazing: The moment of "interested" is defined as a moment in which the wearer of a camera is gazing at something, e.g., people, objects, landscape. We consider that people who have interests in something tend to keep their body still to fix their eyes and thus blurring is reduced.

Similar to VAL features, the degree of gazing (DoG) is defined by two aspects: absolute value and relative one, which are represented A-DoG and R-DoG, respectively. A-DoG is calculated by the sum of inter-frame histogram differences within one second, which means that smaller value indicates less movement and thus more interested. Here, gray-scale images are utilized. The difference of consecutive DoG values are used to obtain R-DoG values. By changing the time difference between frames, i.e., 250 ms and 500 ms, and gray-levels, i.e., 8, 64, and 256, a total of 12 features are calculated.

3.3 Implementation

Data Collection: Two estimators are built using supervised machine learning technique, which requires labeled datasets. We collected datasets of five events from two to six persons including the wearer of the video camera. The summed duration of the events is about 160 min. Table 1 summarizes the datasets. The audio and video is captured by Panasonic A1H [8] at 60 fps, and analyzed using OpenCV, in which an audio channel is separated. The audio channel was originally sampled at 48 kHz; however, the collected data were later down-sampled at 8 kHz to reduce the amount of data.

Table 1. Events in data collection

Event	Number of participants	Duration [min]
Playing darts	2	50
Playing in an amusement park	6	10
Playing a table game	4	40
Singing in a karaoke room	2	30
Drinking in a bar	3	30

Data Labeling: The labels for "partying" were added for each event by the participants of the event including the wearer of the camera every one second. The label "partying" was assigned if at least one person agreed, and the remaining part of the data was labeled as "others". By contrast, only the wearer labeled for "interested" because the state was wearer-dependent one. Similarly, other periods of time except for "interested" was assigned to "others". In total, 9580 time frames were provided for "partying" estimator, in which 324 time frames were labeled as "partying" and the rest of the dataset (9182) were "others". Regarding "interested" estimator, 292 time frames were labeled as "interested",

while "9288" were "others". Due to the large unbalance in each dataset, we reduced the number of data of "others" to the same number as "partying" or "interested" with random sampling.

Feature Selection and Basic Estimator Performance: In Sect. 3.2, we specified the candidates of estimation features, in which 12 features were defined for both "partying" and "interested" estimators. However, the candidates may include redundant ones that can degrade the classification performance and over-consume processing power. So, we applied correlation-based feature subset evaluation method [4] for each feature groups, i.e., VAL and DoG, in combination with greedy stepwise forward selection of best feature subset. As a result, 11 features were selected from VAL features, while 7 were from DoG features. Table 2 shows the selected features. The results of 10 fold cross validation using RandomForest classifier were 0.810 and 0.731 in F-measures for "partying" and "interested" classification, respectively. Meanwhile, the F-measures of all (12) features were 0.805 and 0.731, respectively. So, the performance of "partying" estimator was slightly improved by selected features. However, the performance of "interested" estimator was not improved, and only one feature was removed.

Table 2. Selected features

For "partying" estimator (11 features)	For "interested" estimator (7 features)
$VAL_{100,5}$, $VAL_{500,5}$, $VAL_{1000,5}$, $VAL_{100,7}$, $VAL_{500,7}$, $VAL_{1000,7}$, R-$VAL_{100,5}$, $VAL_{500,5}$, $VAL_{1000,5}$, R-$VAL_{100,7}$, $VAL_{500,7}$	$DoG_{8,250}$, $DoG_{64,250}$, $DoG_{256,250}$ $DoG_{8,500}$, $DoG_{64,500}$, $DoG_{256,500}$ R-$DoG_{8,250}$

Integrating the Results of Two Estimators: The two binary classifiers judge if a given period of time represents the moments that make people who participated in the event feel pleasurable when they watch the images afterwards. A likelihood is obtained from the output of the binary classifier; we simply define the ratio of trees in RandomForest classifier that voted to "positive" class, i.e., "partying" and "interested", to the total number of trees as likelihood in the prototype implementation.

A single score, i.e., fine picture score (FPS), is obtained by weighted averaging. Here, the scores from the two classifiers are equally weighted as defined by Formula (1). To select output images, the system judges if the given moment should be accepted or rejected by applying a specific threshold against the fine picture score. In this paper, we do not apply thresholding. Instead, we investigate the relationship between subjects' ratings and the fine picture scores in next section.

$$FPS = 0.5 \times likelihood(\text{“}partying\text{”}) + 0.5 \times likelihood(\text{“}interested\text{”}) \quad (1)$$

4 Experiment

To understand the emotional effects by the system generated images, a user study was carried out.

4.1 Methodology

Three groups of three students participated to three different types of activities (Table 3). The group members know each other. One subject for each group wore the device and shot the video of the event, and another person became a wearer in another events. One event takes 15 to 20 mins.

One week after shooting, an interview session was held, where the participants rated 90 images (30 images × 3 events) from 1 (do not want to put it to their album at all) to 5 (definitely want to put it to their albums). The system calculated fine picture scores. In rating, pictures with various "fine picture score" were randomly selected and presented to the subjects in a random order. Note that the subjects did not know the fine picture scores that they were rating. Additionally, each wearer was asked to label the movie to either "partying", "interested", and "others" every one second to evaluate the accuracies of the two estimators.

Table 3. Different events performed in user study

Event	Characteristics
Walking	May not be watching conversation partner in talking
	Frequent and unstable gaze movement
	Shooting outdoors
Conversation at a table	My be watching conversation partner in talking
	Infrequent and unstable gaze movement
	Shooting indoors
Playing a table game	May not be watching conversation partner in talking
	Infrequent and periodic gaze movement
	Shooting indoors

4.2 Result

Impression on Pictures with Various Fine Picture Scores: Figure 2 shows the examples of pictures that have different levels of fine picture scores and user ratings. The pictures around the diagonal line from (0.0, 0.0) to (1.0, 1.0) indicate that the system's judgements are close to the subjects' feelings. Pictures B, E, and F have high subjects' ratings as supported by the subjects who argued that people in these pictures looked enjoying, although fine picture score of B is 0.00. By contrast, D and G have low user ratings because G was taken

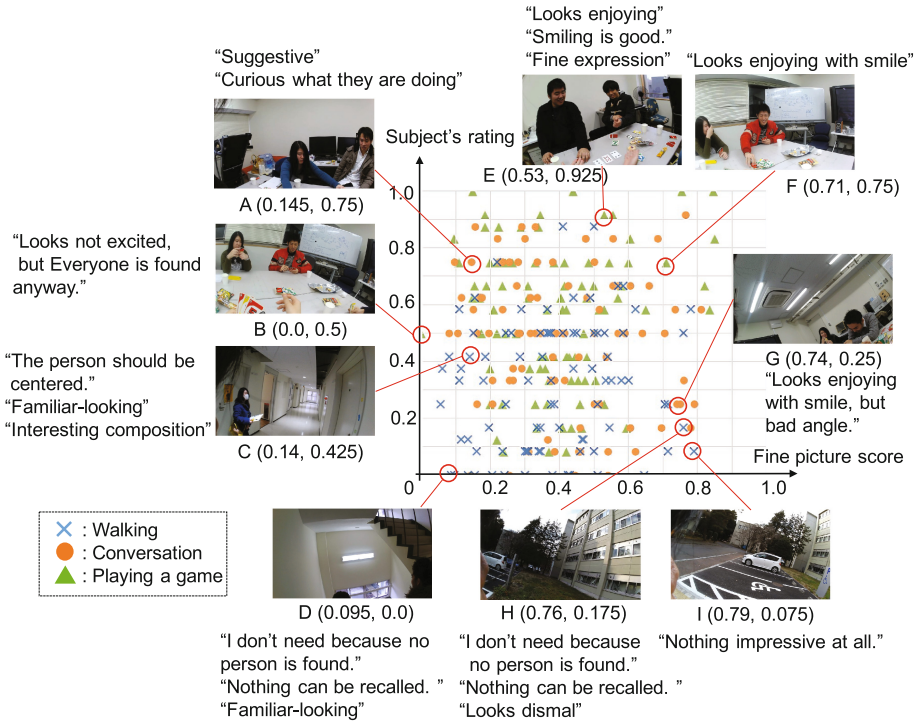

Fig. 2. Pictures with various pairs of fine picture score (the first element in the parentheses) and user rating (the second element). The subjects' comments are shown near each picture.

at too-much upper angle (bad angle) and no person is found (not impressive at all) in D although G has high system score (0.74).

Average correlation coefficients between system's judgements and subjects ratings per event are summarized in Table 4, which are calculated against the likelihood value of "partying" classification, "interested" classification, and fine picture score. The correlation coefficient of fine picture score does not show high correlation, i.e., the value for "walking" event (0.105) is the highest, and negative correlation exists in "conversation" event (−0.088). Regarding the likelihood of "partying" classifier, the value of "conversation" shows the highest (0.197), but negative correlation exists in "walking". Meanwhile, positive correlation is found only in "walking" (0.161) in the likelihood of "interested".

Figure 3 shows the breakdown of subjects' ratings per event category. One-way ANOVA shows that significant difference exists in the event types ($F(2, 267) = 3.03$, $p < 0.05$). The figure indicates that about half of the pictures in the "walking" category got negative impression of rating 1 or 2, while almost half of pictures in "gaming" had positive impression of rating 4 or 5. Opinions against pictures that all the subjects rated highest score "5" are "interesting

moment was shot", "I can imagine their pleasurable moment as well as what they were doing", "I can understand the serious situation." By contrast, the pictures that had lowest score "1" were said that "none appears", "very blurred picture", "nothing impressive at all". Other comments from the subjects and the pairs of fine picture score and the subjects' ratings are mapped in Fig. 2.

Table 4. Correlation of system's judgements and subjects' ratings

	Walking	Conversation	Gaming
Likelihood of "Partying"	−0.023	0.197	0.167
Likelihood of "Interested"	0.161	−0.149	−0.097
Fine picture score	0.105	−0.088	0.077

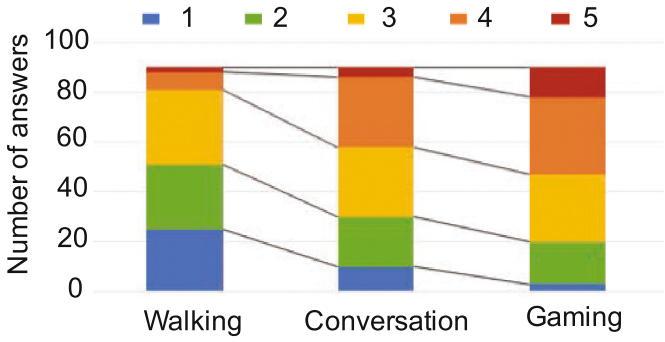

Fig. 3. The breakdown of subjects' ratings per event category. Rating 1 means that a subject does not want to keep the picture into his/her album, while he/she definitely do so in case of rating 5.

Accuracies of Two Estimators: Figure 4 shows box plots that show the difference in likelihood depending on the presence of labels from the subjects. In the figure, "Labeled" indicates that the subjects considered that the period of time was about (a) "partying" and (b) "interested", respectively. By contrast, "Not Labeled" means that no label was put by the subjects. Note that a period of time was labeled "partying" or "interested" if at least one subject agreed on. T-tests show that significant differences exist between "Labeled" and "Non Labeled" in the average of likelihood ($t(391) = 1.26 \times 10^{-76}$, $p < 0.05$ for "partying" and $t(7944) = 2.01 \times 10^{-4}$, $p < 0.05$ for "interested"). As shown in (a), there is a large difference depending on the presence of the label of "partying", which suggests that the system's estimation on the moment of "partying" is closer to the subjects' decision criteria than that of "interested".

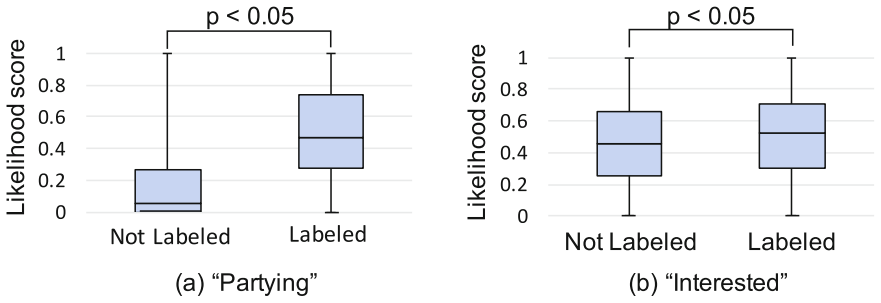

Fig. 4. Box plots of likelihood from user labeling.

5 Discussion

5.1 Factors that Affect Subjects' Rating of Pleasurable Moment

As shown in Fig. 2 and Table 4, little correlation is observed between fine shot scores and the subjects' decisions. This means that the system's decisions are not always consistent with the subjects' impressions. For example, picture G in Fig. 2 has high fine picture score (0.74), while the subject's rating is low (0.25) because the angle of the picture is not fine. Interview results reveal factors that degrades subjects' ratings as follows:

- High predictability and familiar-looking of the situation
- Small number of clues in the pictures
- Awkward angle and ill-composition of pictures
- Blurring pictures
- Small number of person in the picture
- Negative facial expression
- Duplication of pictures that are already presented

Subjects rated high score on the pictures that remind the subjects of the situations where a person who rarely tells a joke happened to do so and a weak game player accidentally won a game. This is because the situation is not predictable. By contrast, mundane situations failed in reminding the subjects of any special emotion and thus had low ratings (C, D). In addition to predictability, the number of clues in the picture may affect rating. Pictures H and I had opinions that they were not recallable and impressive, while a subject insisted that the deal in picture B reminded him of an exciting moment even though they were not smiling. These factors are content-dependent, which might be realized by image and audio understanding and person profiling based on big-data analysis.

As content-independent factors, awkward angle of pictures (G), ill-composed (C) and blurring pictures had low ratings. Duplication of pictures should also be avoided; actually, only the first pictures gained high ratings even though similar pictures existed. To remove such irrelevant pictures, blurry image detection [2] as well as keyframe selection [2,9] should be applied. Subjects liked pictures with

more friends and fine facial expressions (B, E, F), rather than scenes without persons (D, H), which means the number of people and their facial expressions are important factors.

5.2 Difference in the Type of Engaged Events

As shown in Sect. 4.2, there is significant difference in the type of events. The "walking" category had lowest average rating. We consider that this comes from misidentification of "interested" moment. The periods of walking as his/her faces forward and standing still for a break might be judged as "interested" because the moment is determined based on the stability of frames in a video and very few camera fluctuation is observed during the period. Therefore, extra features that filter out such situation, i.e., the camera is stable but the user is not interested in anything, should be investigated. Other reasons could be because walking on the same old way did not provide any extraordinary scene and seeing friends from behind did not remind the subjects of any special emotion.

The event "gaming" obtained the highest average rating, which suggests that the subjects did not only remember pleasurable moments from the smiles in the picture, e.g., pictures E, F, and G in Fig. 2, but also the number of clues in the picture to recall the situation may affect rating as discussed in Sect. 5.1.

5.3 Integrating Fine Picture Scores of Two Aspects

Table 4 implies that an aspect of "interested" reflects the subjects' feelings in a situation with motion, e.g., "waling" although the correlation was very small. By contrast, an aspect of "partying" seems to fit the subjects' feelings in a situation with limited motion such as "conversation". We consider that this is because the gaze detected at a scene where it is always moving is more important than gaze detected at a scene with less motion. In addition, the degree of gaze in "conversation" or "gaming" tends to be always large, which makes it difficult to detect the moment that people are truly gazing with interest. Therefore, it is suggested that the influence of the aspect of "interested" becomes large in a situation where camera wearers move a lot, and that of "partying" increases in a situation with little movement. In this paper, the fine picture score is calculated with equally-weighted average of likelihood of these two aspects (see Formula (1)); however, the final score will be improved by changing the weight depending on the situations.

6 Conclusion

We proposed a system to extract still images from a first-person viewpoint video taken during a group activities to allow the members to recall pleasurable moments. The audio and video features that characterize the states of "partying" and "interested" are defined. Through a preliminary user study, we found content-independent factors that affect the likeability of the output images, as

well as factors that need deep-understanding of the events in the picture and person profiling.

We are planning to enhance the system by dealing with content-independent factors such as the number of people, facial expression, angle, composition, and uniqueness of the moment.

Acknowledgment. This work was partially supported by a JSPS Grant-in-Aid for Scientific Research: 15K-00265.

References

1. Bolanos, M., Dimiccoli, M., Radeva, P.: Toward storytelling from visual lifelogging: an overview. IEEE Trans. Hum. Mach. Syst., 1–14 (2017). http://ieeexplore.ieee.org/document/7723826/
2. Chowdhury, S., McParlane, P.J., Ferdous, M.S., Jose, J.: "My day in review": visually summarising noisy lifelog data. In: Proceedings of the 5th ACM on International Conference on Multimedia Retrieval, ICMR 2015, pp. 607–610. ACM, New York (2015). http://doi.acm.org/10.1145/2671188.2749393
3. GoPro. Inc.: GoPro. http://gopro.com
4. Hall, M.A.: Correlation-based feature selection for machine learning. Ph.D. thesis, The University of Waikato (1999)
5. Healey, J., Picard, W., R.: StartleCam: a cybernetic wearable camera. In: Proceedings of the 2nd IEEE International Symposium on Wearable Computers, ISWC 1998, pp. 42–49 (1998)
6. 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). http://link.springer.com/10.1007/11853565_11
7. Jinda-Apiraksa, A., Machajdik, J., Sablatnig, R.: A keyframe selection of lifelog image sequences. In: Proceedings of IAPR International Conference on Machine Vision Applications, Kyoto, pp. 33–36 (2013). http://www.mva-org.jp/Proceedings/2013USB/papers/03-04.pdf
8. Panasonic Corp.: A1H wearable camera. http://panasonic.jp/wearable/a1h/
9. Ratsamee, P., Mae, Y., Jinda-apiraksa, A., Horade, M., Kamiyama, K., Kojima, M., Arai, T.: Keyframe selection framework based on visual and excitement features for lifelog image sequences. Int. J. Soc. Robot. **7**(5), 859–874 (2015)
10. Sellen, A.J., Fogg, A., Aitken, M., Hodges, S., Rother, C., Wood, K.: Do life-logging technologies support memory for the past?: an experimental study using sensecam. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI 2007, pp. 81–90 (2007). http://dl.acm.org/citation.cfm?id=1240636