2013 IEEE/RSJ International Conference on
Intelligent Robots and Systems (IROS)
November 3-7, 2013. Tokyo, Japan

Lifelogging Keyframe Selection using Image Quality Measurements and Physiological Excitement Features

Photchara Ratsamee[1], Yasushi Mae[1], Amornched Jinda-apiraksa[2], Jana Machajdik[3],
Kenichi Ohara[4], Masaru Kojima[1], Robert Sablatnig[3], and Tatsuo Arai[1]

***Abstract*— Keyframe selection is the process of finding a representative frame in an image sequence. Although mostly known from video processing, keyframe selection faces new challenges in the lifelog domain. To obtain a keyframe that is close to a user-selected frame, we propose a keyframe selection method based on image quality measurements and excitement features. Image quality measurements such as contrast, color variance, sharpness, noise and saliency are used to filter high quality images. However, high quality images are not necessarily keyframes because humans also use emotions in the selection process. In this study, we employ a biosensor to measure the excitement of humans. In previous investigation, keyframe selection using only image quality measurements yielded an acceptance rate of** 79.70%**. Our proposed method achieves an acceptance rate of** 84.45%**.**

I. Introduction

Lifelogging [1] refers to recording daily life using multiple wearable sensors such as a camera to capture images, a microphone to record conversations and surrounding sounds, GPS to track positions, and so on. Currently, the research on lifelog image processing or visual lifelogging is becoming more active since it has many applications in medical, touristic or human attention analysis. Since lifelog image sequences contain many images, the most important and meaningful frame (keyframe) of each event should be properly selected. A keyframe can be one of the frames in the sequence, or multiple images combined into one summary image [2]. In this study, we are interested in the selection of only one keyframe from each image sequence in the event.

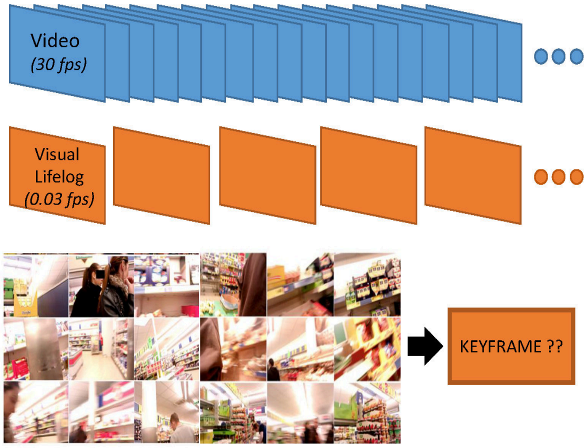

Fig. 1. Visualization of image obtained from video sequence and lifelogging devices. There are obvious differences in frame rate between both image domains. Noncontinuous image sequences are obtained from a lifelogging device. This kind of image sequence poses difficulty as far as event segmentation and keyframe selection using related features between consecutive frames are concerned.

Keyframe selection was first used in video processing as a method of selecting a thumbnail picture. The conventional method selects a keyframe from a fixed position, such as the middle frame of the sequence [3] for simplicity and fast processing time. In other methods, the keyframe is selected based on visual criteria such as motion [4] or the presence of humans in the image. In comparison with video processing, where frames are usually captured at 24-30 *fps*, images from a visual lifelog device are passively captured in larger discrete time intervals, e.g. 1 frame every 30 seconds (0.033 *fps*) to record daily life activity. Due to the low frame rate, there are great differences in consecutive images in lifelog image sequences (as seen in Fig. 1). The keyframe selection technique, which benefit from the properties of consecutive frames [4], [3] can not be applied directly in the lifelog domain.

Visual lifelogging devices are widely available nowadays, for example Sensecam [5]. Such a device enables scenarios such as event capture, story-telling, and mem-

*This work was supported in part by the Grant-in-Aid for Scientific Research (C) 23500242.

**This work was done when Phochara Ratsamee was a visiting student at Computer Vision Lab, Vienna University of Technology (TU Wien), Austria.

[1] P. Ratsamee, Y. Mae, M. Kojima, and T. Arai are with Faculty of Systems Science and Applied Informatics at Osaka University, Japan. E-mail: {ratsamee, mae, kojima, arai}@arai-lab.sys.es.osaka-u.ac.jp

[2] A. Jinda-Apiraksa is at Advanced Digital Sciences Center (ADSC), University of Illinois at Urbana-Champaign, Singapore. He is now supported by the research grant for ADSC's Human Sixth Sense Programme from Singapore's Agency for Science, Technology and Research (A*STAR). E-mail: amornched.ja@adsc.com.sg

[3] Jana M. and Robert S. are with Computer Vision Lab, Institute of Computer Aided Automation, Vienna University of Technology (TU Wien), Austria. E-mail: {jana, sab}@caa.tuwien.ac.at

[4] K. Ohara is with Faculty of Science and Technology, Meijo University, Japan. E-mail: kohara@meijo-u.ac.jp

978-1-4673-6357-0/13/$31.00 ©2013 European Union

ory assistance. Doherty *et al.* [6] proposed a keyframe selection technique for lifelogging based on many wearable sensors and image features. For evaluation, the method is compared with the middle frame selection method and is found to yield improved results. They show that each sensor has unique benefits in each particular situations. Amornched *et al.* [7] proved that this technique is robust to high variability in passively captured image collections. They also proposed a keyframe evaluation framework to evaluate the degree of representativeness of the keyframe and discussed the nature of position distribution of the user selected keyframe choices.

Previous keyframe selection techniques rely mainly on image quality measurements and multi-sensor fusion. However, there still remain unsolved problems in the domain of processing passively captured images. One of the main reasons is that keyframe selection is a highly subjective issue. From our preliminary experiment when users select the keyframes by themselves, they do not consider only image quality, but their emotions are also involved in the selection process. The emotional components are difficult to realize using only vision techniques.

The objective of this study is to automatically select keyframes for all lifelog events that closely match the image chosen by the user. Our assumption is that the most important or memorable moments of human life are those where humans are emotionally involved. We propose a keyframe selection method based not only on image quality but also emotional criteria. We detect excitement, which is one of the emotional criteria, by utilizing wearable bio-sensors [8] that the user will wear alongside the camera. The wearable biosensor quantifies emotional excitement by measuring physiological responses in skin conductance. Images that satisfy both image quality and emotional criteria will be selected as keyframes.

This paper is structured as follows: section II presents our proposed keyframe selection using image quality measurements and the excitement feature. We evaluate our proposed method in Section III. Section IV describes the experiments and results of keyframe selection. Finally, our conclusion and future work are discussed in Section V.

II. Methodology

When a user manually selects the keyframe from the enormous amount of images in each event, the user usually picks the image that has a good quality and is meaningful for him/her [7]. Therefore, we designed a method for keyframe selection based on image quality measurements and excitement features. The structure of the proposed keyframe selection is shown in Fig. 2.

A. Image Quality Measurements

First, we consider the image quality. To measure the quality of each image, the following 5 measurements based on [6], [7] are considered:

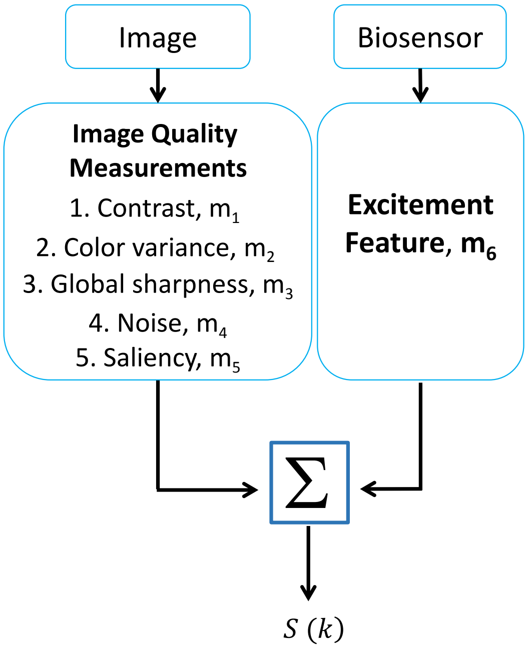

Fig. 2. The structure of the proposed keyframe selection

1) *Contrast (m_1):* The human eye is sensitive to images that have optimal contrast [9].
2) *Color Variance (m_2):* High color variation in the image strongly relates to the degree of colorfulness [6].
3) *Global Sharpness (m_3):* This feature is used to filter blurry images. The method is implemented according to [10].
4) *Noise (m_4):* Heavily noisy images are rarely preferred by humans. We measure the noise level as described in [7].
5) *Saliency (m_5):* Saliency measurement [11], [7] is used to measure the degree of occlusion in images.

B. Excitement Feature (m_6)

To detect human excitement, we use the Q sensor by *Affectiva* [12]. This device detects Electrodermal Activity (EDA). EDA represents the change in the conductivity of electricity in the human skin, which increases in cases of emotional arousal, increased cognitive workload and physical exertion. A low level of EDA can refer to the situation when a human is inactive or relaxed. However, EDA offers no indication about the valence of the emotional state (i.e. it does not distinguish between being peacefully calm or bored, or between joyous or angry). For preliminary observations, we investigated the EDA data when users traveled to new places. During the highlight moment of the day such as walking through the castle and souvenir shop (Fig. 3), the EDA value increased significantly. The keyframes selected by users were the ones where the EDA value was also relatively high.

To process the EDA signal, we first filter the raw data

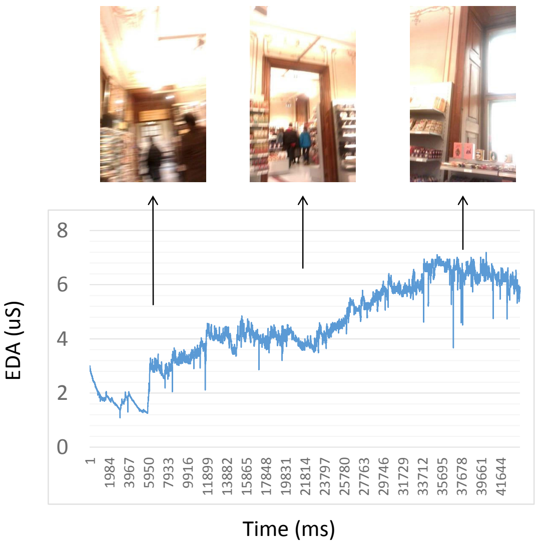

Fig. 3. An example of biosensor data corresponding to images captured by visual lifelogging device

to remove motion artifacts and noise (that appear as high-frequency spikes). A butterworth low-pass filter was applied as follows:

$$|H(f)| = \frac{H_0}{\sqrt{1+(f/f_b)^{2n}}}, \tag{1}$$

where n is the order of the filter, f_b is the cutoff frequency and H_0 is the gain magnitude. Note that the signal is more accurate when the sensors warm to body temperature, and perspiration appears at the contact interface. We normalized the data and then map with the corresponding frames.

C. Keyframe Selection

Since there are 5 image quality measurements and 1 excitement feature involved in making decisions on the keyframe, the super position principle is used to integrate all the features. A weighting factor, w_i, is introduced to each measurement and feature. Hence, the image score (S(k)) is computed from the image features ($m_1 - m_5$), and the bio-physical excitement feature (m_6) of each image at each sampling time is computed by

$$S(k) = \sum_{i=1}^{6} \frac{w_i m_i(k)}{\eta_i}, \tag{2}$$

where $m_i(k)$ is the measurement or feature i of frame k, and η_i is the variance of each feature i in each event used for normalization.

This technique can be extended easily when the number of sensors increases. After calculating the scores of all images in the event, the frames with the highest and most distinct score [7] will be chosen, as presented in Fig. 4.

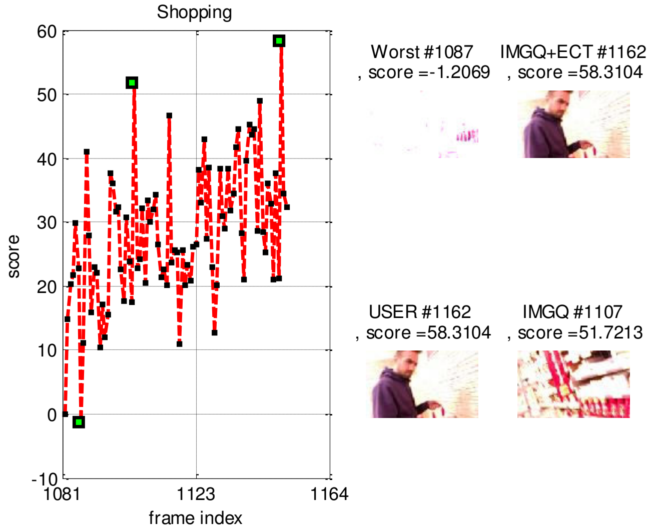

Fig. 4. The graph of scores for each image in the event. The image that got the worst score ('Worst'), highest and most distinct score ('IMGQ+ECT'), the middle image ('MID') and the user selection ('USER') score are also presented.

III. Evaluation Methods

In this evaluation section, the following terms are used to describe the utilized methods. Our proposed keyframe selection (using a combination of image quality measurements and excitement features) is referred to as 'IMGQ+ECT'. Keyframe selection using only image quality measurements is referred to as 'IMGQ'. As a baseline method, we use middle frame keyframe selection, which we abbreviate as 'MID'. Finally, the reference in this study is the keyframe manually selected by the users themselves, named 'USER'.

To evaluate the accuracy of each method, the keyframe selection results ('IMGQ+ECT','IMGQ','MID') and the keyframes from the user selection ('USER'), are compared. In the direct comparison method, only the exact same image as the one chosen by the user should be considered a positive result. However, there are many similar images in the dataset, and therefore picking an image that is similar to the one chosen by the user is also an acceptable result. Therefore, we use the evaluation framework based on the similarity criteria described in [7].

There are 4 similarity criteria: the number of Speeded Up Robust Features (SURF) [13] matching points, the average of the SURF matching error,the color histogram intersection, and the frame number distance.

1) *SURF matching points :* High SURF matching points are directly connected to the similarity between images.
2) *SURF matching error :* Matching error is the degree of dissimilarity.
3) *Color histogram intersection :* Apart from the matching points, similarity in color distribution is also considered.

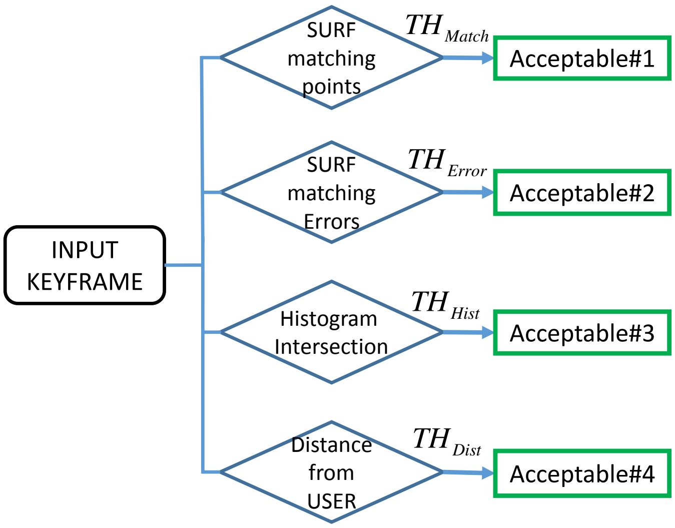

Fig. 5. The evaluation process

4) *Frame Distance :* Frames that are close to the user selected frame have a higher chance to contain similar contents or similar scenes than the frames that are further away.

A flow chart showing how each criterion is used to evaluate the quality of the keyframe is presented in Fig. 5. Each criterion has its own threshold value, i.e., TH_{Match}, TH_{Error}, TH_{Col} and TH_{Dist}, corresponding to SURF matching points, SURF matching errors, histogram intersection, and distance from the manually selected ground truth frame, respectively.

With this evaluation process, the keyframes obtained with the proposed method can be classified as:

- *'Exactly Matched'* refers to the result of keyframe selection that matches perfectly with the USER keyframe.
- *'Acceptable'* is when the keyframe from the proposed method is not exactly the same as the USER keyframe, but it satisfies the similarity conditions in the evaluation process. In other words, it is acceptable as a keyframe. Unlike [7], we modify the diagram in favor of acceptable keyframe analysis. The more criteria the keyframe satisfies, the higher its acceptance score. If the keyframe satisfies all criteria, it is considered as 100 % acceptable. Otherwise, it is considered as 75 %, 50 % and 25 % acceptable when it satisfies 3, 2, and 1 criterion, respectively.
- *'Unacceptable'* refers the keyframe that does not satisfy any similarity criteria.

A. Parameter Optimization

In both the keyframe selection and the evaluation method, there are several parameters that have to be optimized, namely, the weighting factor (w_1 - w_6) and the threshold values of the evaluation process. We follow the parameter optimization process from [7], [14]. In brief, to find the optimal value for all the parameters, we first initialize the weighting factors for the keyframe selection process with equal weights. Then, we use those weighting factors to find proper threshold values in the evaluation process. Finally, we use the average accuracy of each factor to recalculate the weighting factors in the keyframe selection.

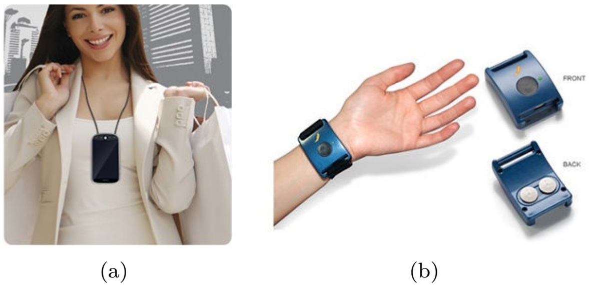

Fig. 6. (a) The Emocam camera on the smartphone used in this study (Image obtained from www.clingo.com); (b) The biosensor worn around wrist for measuring EDA. Picture is taken from [15].

IV. Experiments and Results

A. Experiment Setup

Our study was conducted with 6 participants, who have no prior knowledge about this project, wearing an Emocam (an Android phone with a customized version of the Ubiqlog Application) and bioseosor for a certain amount of time (3-4 hours) during their daily life, as presented in Fig. 6. Datasets from different participants are recorded over a time period of 2-3 weeks. There are 25,451 images in 253 log events. Events range from daily life activities such as using a computer, watching TV, or shopping, to more extraordinary ones such as traveling and sightseeing. A sample of lifelog images is shown in Fig. 1. The lifelog image collection is a mixture of high and low quality images. The implemented keyframe selection and evaluation method run on a PC (E5420 2.50 GHz Xeon CPU, 4096M RAM, NVIDIA Quadro FX 1700 graphic card). The processing time of each frame and the evaluation process varies between 10 ms to 25 ms, depending on the number of SURF keypoints found in the image stream. All data processing is performed using MATLAB. The images from Emocam are time-synchronized with the biosensor to make sure that it records the correct corresponding information.

B. Experiment Results

This section presents the keyframe selection result of each method. The results presented in the Table I are the percentages of *acceptable* keyframes from our keyframe evaluation framework. The preliminary experiments on MID and IMGQ keyframe selection are used to analyze the performance of keyframe selection when compared to our proposed IMGQ+ECT method.

Based on the analysis in [7], most of the USER keyframe locations are close to the middle frame of the

TABLE I

The accuracy results of the proposed keyframe selection algorithm

Keyframe selection method	Acceptance rate (%)	The number of exact keyframes
IMGQ+ECT *(training weighting factors)*	84.45	52 (out of 253)
IMGQ+ECT *(equal weighting factors)*	83.33	49 (out of 253)
IMGQ *(training weighting factors)*	79.70	36 (out of 253)
IMGQ *(equal weighting factors)*	78.11	32 (out of 253)
MID *(based line method)*	67.12	11 (out of 253)

events. The middle frame method is chosen because it is fast and simple. However, our result shows that strictly selecting the middle frame as a keyframe achieves only 67.12 % of acceptable keyframes. This is because the MID keyframe selection method has no guarantee of returning a high quality image. Therefore low quality images (e.g. blurry, or with low contrast) might appear as keyframes.

We firstly investigate the USER keyframe using IMGQ keyframe selection. We found that USER keyframes tend to have a high visual quality. However, the number of acceptable keyframes and exactly matched keyframes is still low. This is because IMGQ keyframe selection method sometimes returns a high quality image, but not an image that is meaningful for the user. As shown in Fig. 7, although the IMGQ method gives a high quality keyframe (a picture of a tree and sky), the user decides to select as a keyframe an image that contains people/activity/places that are meaningful or impressive for his/her, such as his friends smiling or the main building seen while sightseeing.

By integrating the excitement feature from the biosensor data, our proposed method outperforms the result from the MID and IMGQ keyframe selection methods. Our proposed method achieves 84.45 % acceptance, while the other two methods achieve 79.70 % and 67.12 % respectively. Furthermore, there is a higher number of perfect keyframes obtained with our proposed method (53 out of 253 frames) than with the IMGQ (41 out of 253 frames) and MID method (11 out of 253 frames).

C. Discussion

We have analyzed the acceptable keyframe from each method in comparison with the USER keyframe. With the similarity criteria we used in the evaluation process, each acceptable keyframe can be graded with 4 levels depending on the number of criteria it satisfies. Fig. 8 shows the comparison of the representative percentage of each method. The amount of acceptable keyframes that pass all criteria is clearly higher with the IMGQ+ECT method, reaching more than 60% when compared to IMGQ and MID, which achieve only 34.1 % and 25.5 %, respectively. The number of unacceptable keyframes is also lower with our proposed IMGQ+ECT method (from 31.2 % in IMGQ to 15.5 %).

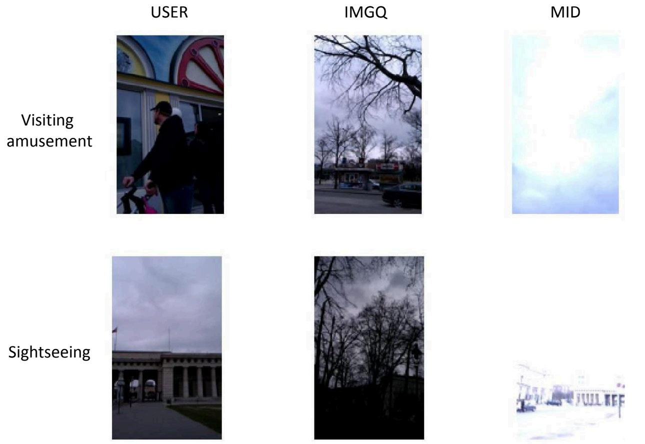

Fig. 7. The result of keyframe selection from IMGQ and MID method. The keyframe selection results are still far from the USER keyframe.

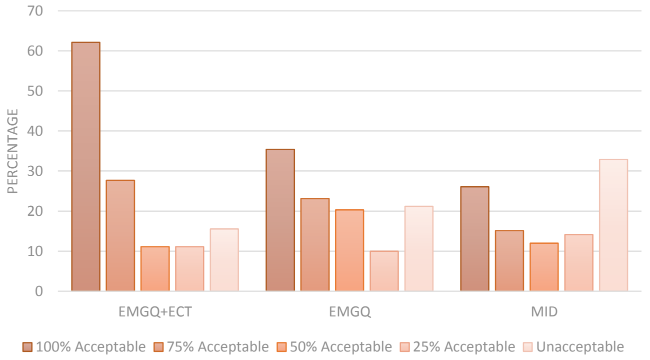

Fig. 8. Percentages of acceptable keyframes; a comparison between the IMGQ+ECT, IMGQ and MID keyframe selection results.

The quality of the proposed keyframe selection result can also be measured by the frame index distance from the user selected keyframe. Fig. 9 shows the keyframes selected with all 3 methods and the distance between these frames and the one selected by the USER. In this case, the distance was normalized to be in the range between 0−1. The distances between the USER keyframe and the keyframes obtained with IMGQ+ECT, IMGQ and MID keyframe are 0.31, 0.53 and 0.75 respectively. The conventional MID keyframe selection has a relatively big distance from the USER keyframe and also has no significant detail or visual similarity related to the USER keyframe. The keyframe from our proposed IMGQ+ECT method is significantly closer to USER keyframe compared to the IMGQ and MID keyframes. Finally, although the IMGQ+ECT keyframe does not exactly match the USER keyframe, it has a very similar content when compared to it.

The proposed method has been developed based on the assumption that humans select a keyframe based on its

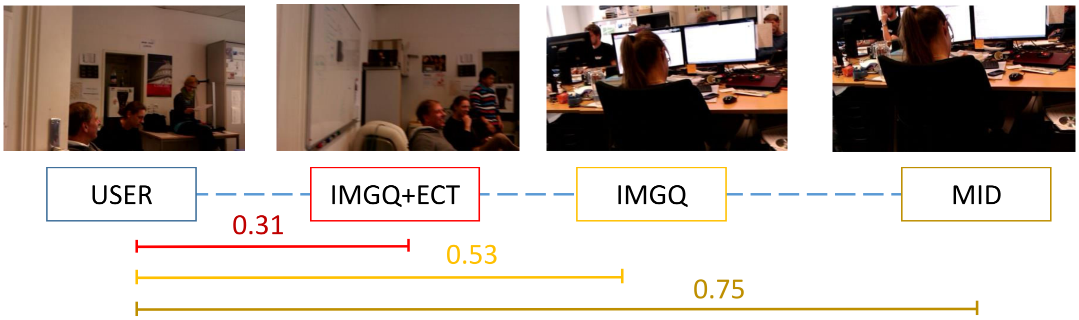

Fig. 9. The comparison of distances from the USER keyframe and IMGQ+ECT, IMGQ and MID keyframe selection results. IMGQ+ECT keyframe selection result is the closest keyframe to the USER keyframe.

quality, as well as their emotion, especially excitement. In case there is no excitement, such as when the user is relaxing or studying, our system will consider other factors involved in keyframe selection e.g. image quality. There are still many factors that have to be implemented in our framework, for example, the presence of a smiling face in the frame; also, adding other sensors to integrate other human senses such as sound or touch must be considered in the future. These other factors can be easily added to our proposed model. Moreover, the parameters in our system were optimized by a simple methodology, and calibrated from our set of participants only. For other groups of people in a different culture or context, all parameters and human characteristics have to be reconsidered.

V. Conclusion

We proposed a keyframe selection based on image quality measurements and emotional features. We used 5 different image quality measurements, i.e., contrast, color variance, sharpness, noise and saliency. By taking into account the influence of excitement measured by a biosensor, the acceptance rate increased to 84.40% compared to keyframe selection using only image quality measurement (79.70%) and middle frame method (67.12%). Moreover, the number of keyframes that match the keyframe selected by the user was significantly improved from 4.34% to 20.09%. The result about frame index distance from the USER keyframe also confirms that using a combination of image quality measurements and excitement in keyframe selection delivers results that are closer to the user selected keyframes with similar content.

VI. Acknowledgments

We would like to thank the 6 student volunteers, who helped us with data collection and ground truth labeling and scoring.

References

[1] V. Bush, "As we may think," *Multimedia: From Wagner to Virtual Reality, Norton, New York*, 2001.

[2] A. Rav-Acha, Y. Pritch, and S. Peleg, "Making a long video short: Dynamic video synopsis," in *Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 1. IEEE, 2006, pp. 435–441.

[3] A. F. Smeaton and P. Browne, "A usage study of retrieval modalities for video shot retrieval," *Information Processing Management*, vol. 42, no. 5, pp. 1330 – 1344, 2006. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0306457305001524

[4] W. Wolf, "Key frame selection by motion analysis," in *Proceedings on International Conference on Acoustics, Speech, and Signal Processing*, vol. 2. IEEE, 1996, pp. 1228–1231.

[5] S. Hodges, L. Williams, E. Berry, S. Izadi, J. Srinivasan, A. Butler, G. Smyth, N. Kapur, and K. Wood, "Sensecam: A retrospective memory aid," *Ubiquitous Computing*, pp. 177–193, 2006.

[6] A. Doherty, D. Byrne, A. Smeaton, G. Jones, and M. Hughes, "Investigating keyframe selection methods in the novel domain of passively captured visual lifelogs," in *Proceedings of the international conference on Content-based image and video retrieval*. ACM, 2008, pp. 259–268.

[7] A. Jinda-apiraksa, J. Machajdik, and R. Sablatnig, "A keyframe selection of lifelog image sequences," in *Proceedings of the IAPR Conference on Machine Vision Applications (IAPR MVA)*, Kyoto, Japan, May 2013, pp. 33–36.

[8] J. Wagner, J. Kim, and E. André, "From physiological signals to emotions: Implementing and comparing selected methods for feature extraction and classification," in *International Conference on Multimedia and Expo*. IEEE, 2005, pp. 940–943.

[9] P. G. Barten, "Contrast sensitivity of the human eye and its effects on image quality." SPIE-International Society for Optical Engineering, 1999.

[10] P. Marziliano, F. Dufaux, S. Winkler, and T. Ebrahimi, "A no-reference perceptual blur metric," in *Image Processing. 2002. Proceedings. 2002 International Conference on*, vol. 3. IEEE, 2002, pp. 57–60.

[11] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 11, pp. 1254–1259, 1998.

[12] M.-Z. Poh, N. C. Swenson, and R. W. Picard, "A wearable sensor for unobtrusive, long-term assessment of electrodermal activity," *IEEE Transactions on Biomedical Engineering*, vol. 57, no. 5, pp. 1243–1252, 2010.

[13] H. Bay, T. Tuytelaars, and L. Van Gool, "Surf: Speeded up robust features," *Proceeding of European Conference on Computer Vision*, pp. 404–417, 2006.

[14] A. Jinda-apiraksa, "A keyframe selection of lifelog image sequences," Erasmus Mundus M.Sc. in Visions and Robotics thesis, Vienna University of Technology (TU Wien), June 2012.

[15] "How to measure emotions." *http://www.kurzweilai.net/how-to-measure-emotions*, 2012.