A Keyframe Selection of Lifelog Image Sequences

Amornched Jinda-Apiraksa [*,†], Jana Machajdik [*], and Robert Sablatnig [*]

[*]Institute of Computer Aided Automation, Vienna University of Technology (TU Wien), Austria
{taro, jana, sab}@caa.tuwien.ac.at
[†]Advanced Digital Sciences Center (ADSC), University of Illinois at Urbana-Champaign
amornched.ja@adsc.com.sg

Abstract

Visual lifelogging is a new concept of recording one's daily life by using wearable camera to automatically capture images of one's surroundings. Keyframe selection is a crucial process for summarizing lifelog image sequences. In the visual lifelog domain, images are passively captured in predefined time intervals, e.g. one image every 30 seconds. This results in variable visual quality in the image sequences. Contrary to videos, two consecutive frames are not necessarily similar in lifelog sequences. Thus, video processing techniques are not directly applicable. We propose a keyframe selection technique based on measuring image quality and distance to the middle frame. Based on the proposed evaluation framework, 81.6 % of keyframes selected by our approach are accepted, whereas only 70.4 % are accepted when using the middle frames as keyframes. Additionally, Ground Truth (GT) keyframes are investigated in terms of image quality and the position in time relative to their events. This provides information about their distributions and explain the results.

1 Introduction

Lifelogging is the process of using wearable computers to automatically record aspects of one's surroundings, e.g. images (visual lifelogging), global positions and so on. The lifelogging idea was firstly proposed by Bush [1] in 1945. However, the research in lifelogging for personal purpose as an active research area started in last decade [2]. Images and possibly other surrounding information, e.g. location or accelerometer, are recorded. The main purposes of personal lifelogging are for touristic and medical purposes. Especially for touristic purpose, it clearly enables scenarios that people are excited about such as event capture, story-telling, and memory assistance [3].

This paper aims to present an automatic way of creating a visual diary from lifelog image sequences, i.e. to summarise daily activities. There are two main processes in the lifelog's workflow, as can be seen in Fig. 1, which are event segmentation (group images into coherent collections) and keyframe selection (select the group's representative image). This paper will focus on the automatic *keyframe selection* process, assuming the event segmentation process has been done separately. A semi-automatic evaluation structure is proposed in order to avoid time-consuming human evaluation.

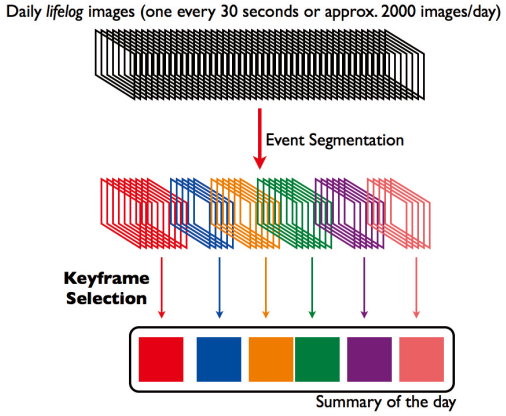


Figure 1. The flow of the lifelong images (i.e. all images are segmented into smaller events and representative image is selected from each event.

Images in the visual lifelogging domain come from passive capturing devices. The lifelogging device captures about 2,000 images per day or approximately 700,000 images every year. For such a large amount of visual information to be useful, it needs to be managed into a proper structure, e.g. by date or event. The keyframe selected to embody each event must be representative of that content and must convey its core concepts. This is a subjective task because images have an inherent underlying semantic meaning [4]. Our framework specifically uses the keyframes from all events of the day to summarise the day, thereby creating a visual diary, as shown in Fig. 1.

Doherty et.al [4] agree that motion analysis, which is a popular mechanism for keyframe selection in video processing, does not necessarily translate directly to visual lifelogs due to the huge difference in frame rate (0.033 and 24 fps.). Instead of calculating the motion from consecutive frames, the motion of the camera itself, which is captured by the accelerometer, is used. In our approach, we remove the frame similarity comparison from Doherty et.al [4] and add the influence of the middle frame in order to reduce the computational time while still maintain similar performance.

This paper has three contributions. Firstly, we propose and prove that the simple quality assessment using the 'nearest neighbour ratio strategy' achieve better results than state-of-the-art, but has much smaller

This work was supported by the European Commission, as part of the of Erasmus Mundus Masters in Computer Vision and Robotics (VIBOT). A. Jinda-Apiraksa 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).

computational complexity. Secondly, we propose the semi-automatic way to evaluate general keyframe selection process based on similarity measurement. This reduces the burden of human evaluation. Lastly, we study two scenarios of keyframe selection, which are the image's quality and position impact, in order to confirm our proposed result.

This paper is organized in the following structure. Section 2 explains our proposed keyframe selection algorithm. Section 3 presents the semi-automatic evaluation framework. Results are shown in Section 4. We also analyze the nature of keyframes in lifelog domain in Section 5. Finally, we conclude our study in Section 6.

2 Keyframe Selection

In our experiments, we recorded 3 different datasets, covering a time period of 3 weeks each. There are a total of 30,926 images in 704 log events. The lifelog image repository is a mixture of high and low quality images. A significant portion (about 40 %) of all images are considered as low quality images, e.g. blurry, under/over-exposure, or occlusion.

Summarizing literature [5, 6, 7], there are three main approaches for keyframe selection: similarity based, feature based, and fixed-position approaches. Due to the ambiguity of assumption and inconvenience of exhausted comparison of the similarity based approach, we decide to use the combination of a feature based and a fixed-position approach. Moreover, a high success rate (average score of 3.92 out of 5 or 78.4 %) of the feature based approach has been proven by Doherty et al. in [4].

2.1 Feature Extraction

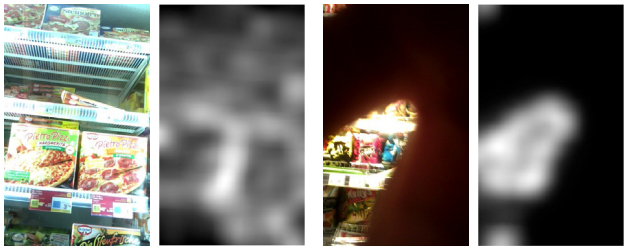
We have extracted 5 low-level image quality measurements based on [4] as described in the following.

1. **Contrast (f_1):** Only Y component image (from YUV color space) is split into 8-by-8 blocks. The sub-contrast of each block is calculated by subtracting the minimum value from the maximum value. The image's contrast is then computed by the mean of all sub-contrast values.

2. **Color Variance (f_2):** Each of the image's pixels is classified into one of the 8 color bins, which are black, white, red, green, blue, yellow, cyan and magenta, based on the smallest Euclidean distance. The color variance of the image is computed by the variance of the number of pixels in all bins that contain more than 20 % of the overall pixels.

3. **Global Sharpness (f_3):** We inherit the technique from Marziliano et al.[8] such that sharp images have thinner edges than blurry images.

4. **Noise Measurement(f_4):** The noise measurement is calculated by the percentage of noisy pixels in each frame. In order to classify each pixel as a noise, the image is divided into smaller 3-by-3 blocks. Then, the distance between each pixel to the block's mean value is calculated. If the middle pixel has the largest distance to the mean compared to all 8 neighbouring pixels, the middle pixel is classified as the *noisy* pixel.



(a) (b) (c) (d)

Figure 2. The demonstration of saliency maps.

5. **Saliency Measure (f_5):** The saliency map calculation from Itti et al. [9] is used. The saliency measurement of the image is the summation of all pixel's saliency values. Saliency measurement is able to solve the occlusion problem. As can be seen in Fig. 2, the non-normalized saliency measure of Fig. 2(b) (normal scene) is 87,267 and the saliency measure of Fig. 2(d) (occlusion scene) is 41,021.

The final quality score is computed by combining all normalized values from the feature extraction, as shown in Eq. 1.

$$QualityScore(k) = \sum_{i=1}^{5} w_i \frac{f_i(k)}{\sigma_i}, \qquad (1)$$

where $f_n(k)$ is the value of feature n of frame k, σ_n is the variance (normalization term) of feature n in each event, w_n is the weighting factor of each feature, and $n = 1, 2, 3, 4, 5$. After applying the same feature extraction to all frames in the event, we can plot the quality score, as shown in Fig. 3.

2.2 Keyframe Selection using the Nearest Neighbour Ratio Strategy

We apply keyframe selection based on the nearest neighbour ratio strategy, similar to [10]. The quality score of the selected images should not only be high, but also distinct. After calculating the scores of all images in the event, two frames with the highest score are chosen. If the ratio of their quality scores is less than 0.7 (similar to the ratio used in [11]), the frame with the highest score will be selected to be the event's keyframe. Otherwise, one of those two frames that is closer to the middle frame will be selected, as illustrated in Fig. 3.

3 Keyframe Evaluation

We propose a framework to evaluate the performance of the keyframe selection algorithm based on the fact that we have only one Ground Truth (GT) keyframe of each event. The ideal result of the keyframe selection algorithm must be exactly the same as the GT keyframe. However, the image with similar contents is also acceptable as a correct selection

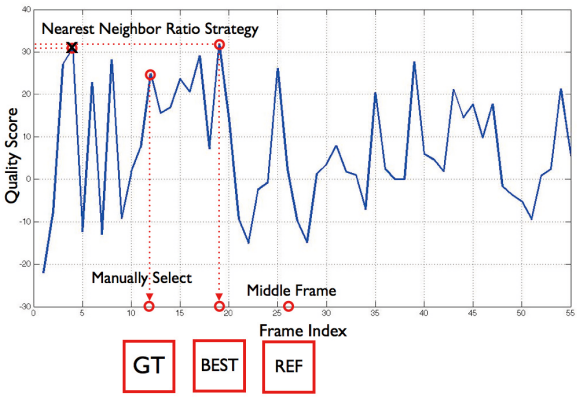


Figure 3. There are 3 generic keyframe terms (GT, BEST, and REF) used in this paper.

because users evaluate the keyframe by its contents. Assume that the GT keyframes are the best frame to represent each event, the proposed framework will evaluate the similarity between the BEST keyframe (proposed algorithm) and the GT keyframe compared to the similarity between the REF keyframe and the GT keyframe of each event. We classify each BEST and REF keyframe as *'accepted'* or *'rejected'* based on their similarity to GT. The *'accepted'* keyframe must have similar contents (e.g. objects or colors).

The proposed framework, as show in Fig. 4, consists of the following 4 similarity criteria :

1. **SURF matching points:** We count the number of SURF (Speeded Up Robust Features) matching pairs based on the nearest neighbour ratio matching strategy, as mentioned in [10].

2. **SURF matching error:** The SURF matching error is calculated using the average of the error from all matching points.

3. **Color histogram intersection:** The *hue* component histograms of both images are compared (hue is used to distinguish colors) and the histogram intersection area is calculated.

4. **Frame Distance:** The frame distance is calculated by the difference of the frame index. It is normalized by the total number of frames in that event.

4 Keyframe Selection Results

We test the accuracy of the proposed keyframe selection algorithm based on our proposed *keyframe evaluation* criteria and compare results with the base line technique, which is selecting the *middle frame*. We also calculate the number of perfect results, which is measured by the number of BEST keyframes which are exactly the same as GT keyframe. All of the results are shown in Table 1.

Both results from the proposed algorithm (80.61 % and 81.63 %) are better than the result of selecting the middle frame (70.41 %) using the same keyframe

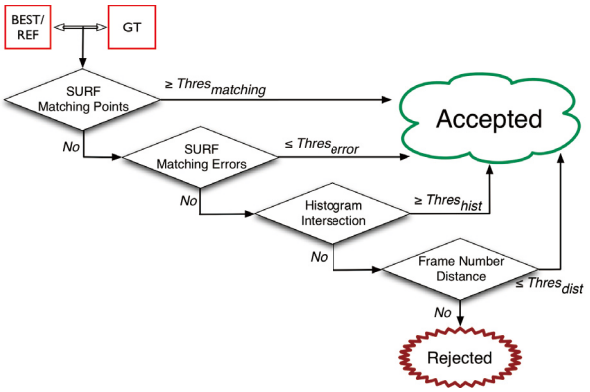


Figure 4. The cascading criteria checking of keyframe evaluation

Table 1. The accuracy results of the proposed and based-line keyframe selection algorithm

Keyframe Selection Method	Accuracy (%)	# of perfect keyframes
Proposed Method		
(equal weight)	80.61	28 *(out of 196)*
(trained weight)	81.63	31 *(out of 196)*
Middle frame	70.41	5 *(out of 196)*

evaluation framework. Note that, those results are the percentage of *accepted* keyframes from our keyframe evaluation framework and do not necessarily represent the event in the best way possible However, there is a higher number of perfect keyframes from our proposed method (31 out of 196 frames) than the base line method (5 out of 196 frames). We have also manually evaluated the performance, as follows, in order to confirm the results from our proposed evaluation framework.

- **How many of keyframes are acceptable as the correct keyframe?**
  There are 83.16 % and 51.53 % of BEST and REF keyframes that are considered as a correct representation of the event respectively. This is a proof that our proposed keyframe selection method outperforms the popular method of selecting the middle frame.

- **How many of BEST keyframes are better to represent the event than REF keyframes and vice versa?**
  For each event, we manually judge whether the BEST or the REF keyframe is better to represent its event than the others. Out of 196 events that we manually evaluate, there are 95 events (48.47 %) where the BEST keyframe is better to represent the event than REF keyframe, while there are only 21 events (10.71 %) vice versa. In other word, there are 175 out of 196 events (89.29 %) where the proposed keyframe selection algorithm perform *better than* or *equal to* the base line method.

5 The nature of the ground truth

This section analyzes the nature of GT keyframes in our *lifelog* dataset. This investigation explains the

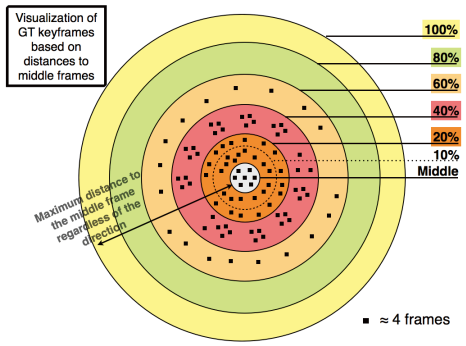


Figure 5. The nature of the GT keyframes distribution based on their location in the event

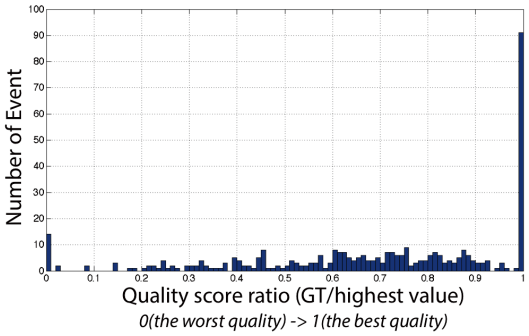


Figure 6. The quality of GT keyframes compared to the highest quality score of their event

performance and suggests future improvements of our keyframe selection algorithm.

Position distribution To analyze the distribution of GT keyframes among the dataset based on their positions, we calculate the frame distance of each GT keyframe to the middle frame of its event. Fig. 5 visualizes the results. The inner most circle represents the closest distance, which is exactly at the middle position. Each of the black squares represent the proportional amount of GT keyframes that have a distance to the middle frame between two indicated values. It is shown that most of the GT keyframe's locations are close to the middle frame of their events. There are 6.81 % of keyframes that locate exactly at the middle of the event. Further investigation shows that 99.21 % of keyframes are located within 50 % of the distance from their center. Hence, selecting the middle frame returns better result than the other positions. However, strict selection of the middle frame does not guarantee a correct result since there is the possibility of picking up an artifact or a meaningless frame. Therefore, a keyframe selection algorithms that is based on, but not strictly to, the distance to the middle frame yield promising performance.

Quality distribution We investigate the quality of the GT keyframe compared to the rest of the frames in the same event. Quality score ratios of the GT keyframes to the best quality frame in each event are calculated and plotted. As shown in Fig. 6, there are 77.17 % of GT keyframes that have the ratio higher than 50 % (up to 23.36 % of keyframes which have the highest quality score). This investigation proves that GT keyframes that were selected by the user, tend to have high visual quality.

6 Conclusion

The main purpose of this paper was to build the keyframe selection algorithm for visual lifelog image sequences. We developed a technique based on the quality measurements (contrast, color variance, sharpness, noise, and saliency measure) and the distance to the middle frame. We also proposed a semi-automatic way to evaluate the accuracy of the keyframe selection technique, which returns similar evaluation trend compared to manual evaluation.

It was shown that our proposed keyframe selection method has significantly better performance than selecting the middle frame method. The results showed that our proposed keyframe selection algorithm has 31.63 % improvement in the accuracy compared to the middle keyframe method using manual evaluation.

We have also investigated the nature of the visual lifelog images. All of the keyframes are located with in 60 % distance from their middle frame. Moreover, their quality measurements are high, i.e. there are 77.17 % of all keyframes that have the quality measurement in the top 50 % of the event. These investigations prove that our selected features (quality measurements) and approach (based on the middle frame) are meaningful and applicable to the real situation.

There are several possible future improvements. For example, increase the number of datatsets or using features based on other sensors (e.g. bio sensor, GPS data) are helpful for understanding the semantic meaning in keyframes.

References

[1] Bush, V., Wang, J.: As we may think. *Atlantic Monthly* **176** (1945) 101–108

[2] Gurrin, C., Smeaton, A.F., Byrne, D., Jones, G.J.F.: An examination of a large visual lifelog. In: *Proc. AIRS*. (2008) 537–542

[3] Gemmell, J., Williams, L., Wood, K., Lueder, R., Bell, G.: Passive capture and ensuing issues for a personal lifetime store. In: Proc. ACM CARPE. (2004) 48–55

[4] Doherty, A.R., Byrne, D., Smeaton, A.F., Jones, G.J., Hughes, M.: Investigating keyframe selection methods in the novel domain of passively captured visual lifelogs. In: Proc. CIVR, New York, USA (2008) 259–268

[5] Vermaak, J., Perez, P., Gangnet, M., Blake, A.: Rapid summarisation and browsing of video sequences. In: BMVC. (2002) 424–433

[6] Dufaux, F.: Key frame selection to represent a video. In: Proc. ICIP. (2000) 275–278

[7] Cooper, M.L., Foote, J.: Discriminative techniques for keyframe selection. In: Proc. ICME, IEEE (2005) 502–505

[8] Marziliano, P., Dufaux, F., Winkler, S., Ebrahimi, T.: A no-reference perceptual blur metric. In: Proc. ICIP. Volume 3. (2002) 57–60

[9] Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. IEEE Trans. PAMI **20** (1998) 1254–1259

[10] Bay, H., Ess, A., Tuytelaars, T., Van Gool, L.: Speeded-up robust features (surf). Comput. Vis. Image Und. **110** (2008) 346–359

[11] Lowe, D.G.: Distinctive image features from scale-invariant keypoints. IJCV **60** (2004) 91–110