35th Annual International Conference of the IEEE EMBS
Osaka, Japan, 3 - 7 July, 2013

Daily Life Event Segmentation for Lifestyle Evaluation Based on Multi-Sensor Data Recorded by a Wearable Device

Zhen Li, Zhiqiang Wei, Wenyan Jia, and Mingui Sun

Abstract— **In order to evaluate people's lifestyle for health maintenance, this paper presents a segmentation method based on multi-sensor data recorded by a wearable computer called eButton. This device is capable of recording more than ten hours of data continuously each day in multimedia forms. Automatic processing of the recorded data is a significant task. We have developed a two-step summarization method to segment large datasets automatically. At the first step, motion sensor signals are utilized to obtain candidate boundaries between different daily activities in the data. Then, visual features are extracted from images to determine final activity boundaries. It was found that some simple signal measures such as the combination of a standard deviation measure of the gyroscope sensor data at the first step and an image HSV histogram feature at the second step produces satisfactory results in automatic daily life event segmentation. This finding was verified by our experimental results.**

I. Introduction

Improvement in lifestyle has received increasing attention by the public in recent years because lifestyle has high significance in health maintenance and disease prevention. If accurate lifestyle data can be acquired from individuals objectively, healthcare professionals will be able to advice and monitor their lifestyle more effectively. For example, lifestyle data can be used to monitor and control calorie intake and expenditure, environmental pollutant exposure, and psychosocial stress. In recent years, lifestyle data collection and analysis have become an emerging research field in biomedical engineering.

Electronic devices and sensors such as camera, audio recorder, and motion sensors (including accelerometer and gyroscope) have been used to record daily life data in free-living individuals. Data captured by these sensors act as an electronic diary which records events and experiences of an individual's daily life. A wrist-worn activity detector containing a 3-axis accelerometer was developed for motion recording [1]. A shoe-based portable physical activity monitor was investigated containing five pressure sensors and a 3-axis accelerometer [2]. Besides these sensors, the gyroscope, which measures orientation, has been used for gait analysis [3]. The SenseCam [4], which is worn around the neck using a lanyard, takes about 3,000 images each day for life experience recording.

In order to capture not only physical activities involving body motions, but also sedentary activities and social events, such as TV watching, eating and social interaction we have investigated a wearable device called eButton [5] embedded with a camera, a gyroscope, an accelerometer, a GPS sensor and a number of other sensors. The eButton is worn in front of the chest in the form of a decorated chest pin [6]. In order to save power and reduce amounts of data, eButton takes one picture (640 pixels by 480 pixels) at a low rate between one and five seconds, adjustable by the user. This device acquires approximately 10,000 pictures (assuming a rate of 4 seconds per picture for 12 hours), along with data recorded by other sensors, and saves them on a micro SD card. The sampling rate for motion sensors is set to 30Hz.

Despite the use of relatively low data rates, it takes a long time to analyze the large dataset if the analysis is performed manually, such as by the user oneself. In order to reduce data analysis burden and help the user recall past activities when the recorded data are examined, an automatic data segmentation algorithm is very helpful. This algorithm compares data epochs sequentially and groups them together when these epochs are similar. As a result, the user does not have to browse thousands of pictures and other data. Instead, he/her only needs to examine a few representative epochs in each data group to identify the segmented event. Because of the high usefulness of this algorithm, it is currently a key component of our data analysis software for lifestyle evaluation.

In the field of image processing, many video and image sequence segmentation methods have been proposed. Most segmentation methods focus on movies or videos. Color and edge features [7], as well as combinations of these and other image features [8], have been used in segmentation algorithms. Motion information, such as motion vectors, has been well studied in image/video segmentation [9]. Mutual information and the joint entropy between frames have been used in shot detection [10]. Many clustering and machine learning methods, such as support vector machine (SVM), have been applied to video boundaries detection [11, 12].

Although image based segmentation algorithms are effective, these algorithms are generally computationally complex. In the wearable system such as the eButton, the location sensor (GPS) and motion sensors including accelerometer sensor and gyroscope are available, providing additional information. Among these sensors, the GPS data are often unavailable in the indoor environment. We thus focus on motion sensors for activity segmentation. Many

*This work was supported by National Institutes of Health Grants No. R01CA165255 and U01HL91736 of the United States, and the National Key Basic Research Program No. 2010CB735903 of China.

Zhen Li is with Dept. of Computer Science and Technology, Ocean University of China, Qingdao, China and Dept. of Neurosurgery, University of Pittsburgh, Pittsburgh, PA, USA (email: lizhen0130@gmail.com)

Zhiqiang Wei is with Dept. of Computer Science and Technology, Ocean University of China, Qingdao, China (email: weizhiqiang@ouc.edu.cn)

Wenyan Jia is with Dept. of Neurosurgery, University of Pittsburgh, Pittsburgh, PA, USA (email: wej6@pitt.edu).

Mingui Sun is with Dept. of Neurosurgery, University of Pittsburgh, Pittsburgh, PA, USA (email: drsun@pitt.edu). Corresponding author.

978-1-4577-0216-7/13/$26.00 ©2013 IEEE

features such as the mean, standard deviation, kurtosis, spectral energy, signal-magnitude area and autoregressive coefficients have been utilized to classify activities [1, 13]. However, most existing algorithms do not target the detection of event boundaries. Although the CombMNZ algorithm is proposed to fuse accelerometer, visual data and light sensor data for segmentation of life log data acquired by the SenseCam [4], it is assumed that the number of segments during each day is fixed, and the computational cost is relatively high.

II. Segmentation Algorithm

A. Overview of Proposed Segmentation Method

In order to segment multi-sensor data quickly and efficiently, we proposed a two-step segmentation method which combines motion and image features as shown in Fig. 1. The first step of this method detects candidate boundaries according to extracted motion features. At the second step, images in the neighborhoods of candidate boundaries are examined for a similarity evaluation. Because the second step is applied to a small portion of the recorded images, our algorithm is significantly faster than the existing ones which require examination of all images in the dataset.

B. Motion Feature Extraction

Since the data from motion sensors provides basic activity information without details, this type of data is ideal for the first-step, coarse-level segmentation. In our device, motion sensors contain an accelerometer and a gyroscope, which are both three-axis sensors capable of characterizing motion in vector forms. Because, in this step, we are more interested in the computational speed than the segmentation precision, simple signal measures such as the standard deviation (STD) within a certain window is calculated from the motion data in each axis. In order to simplify computation further, we sum the STD values for all three axes to form a new feature which is called a S-STD feature. Fig.2 (a) and (b) show typical raw data of the accelerometer and gyroscope, respectively, captured in the same time interval. Fig.2 (c) and Fig.2 (d) show, respectively, the S-STD features of the 3-axis accelerometer and gyroscope data. It can be clearly seen that S-STD features are effective in describing the levels of motion. The S-STD values close to zero represent inactive activities, and large values indicate more active activities.

C. Motion Data Segmentation

In the first step of data processing, all motion data are divided into a set of segments using a threshold value. Three types of segments are defined according to their lengths: short segment (SS), long inactive segment (IS) and long active segment (AS) for the second step of data processing based on image features. The value of threshold T_m is determined experimentally which will be discussed in the experiment section.

The reason why we define these three segments is because different segment lengths are used in the second processing step. For example, SSs are considered insignificant segments which will be merged with long segments. For IS and AS, the degree of motion is an important factor to indicate whether two adjacent segments should be kept or combined. Detailed description will be provided in Section E.

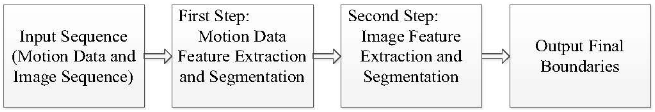
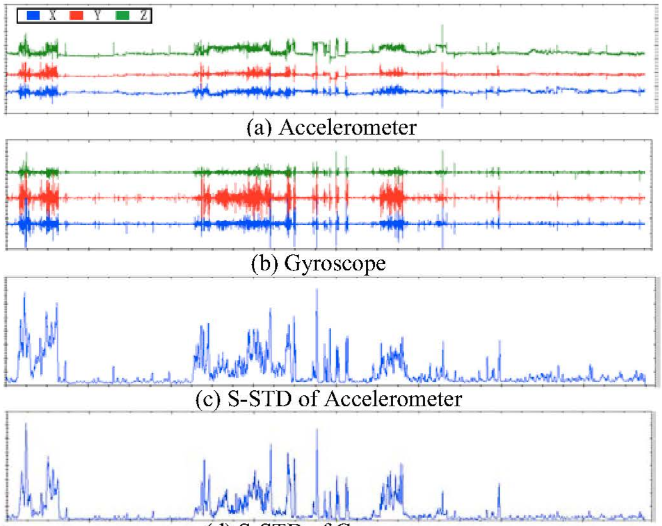

Input Sequence (Motion Data and Image Sequence)
First Step: Motion Data Feature Extraction and Segmentation
Second Step: Image Feature Extraction and Segmentation
Output Final Boundaries

Figure 1. Flow chart of the proposed method



(a) Accelerometer

(b) Gyroscope

(c) S-STD of Accelerometer

(d) S-STD of Gyroscope

Figure 2. Accelerometer and gyroscope data

D. Image Features Extraction

In the second step of data processing, image features are used to fine-tune the boundaries detected in the first step to achieve a higher accuracy. There are many image features to be selected, such as the scale-invariant feature transform (SIFT), HSV histogram, edge histogram, and color layer [14]. Since the frame rate of image data captured by the eButton is very low and the variations between adjacent images are large, many features suitable for a regular frame rate (e.g., 30 frames/second) are not be suitable in our case. We found that global color features could handle the low-rate data more effectively. The color histogram of the HSV channel is one of these features to distinguish different segments. Moreover, the HSV histogram can be calculated more rapidly than other features such as the SIFT and the Edge Histogram Descriptor (EHD). In order to reduce variance in the output, we compute the HSV histogram from a group of images instead of a single image according to $M_i = min\{|S_i|, 100\}$, where $|S_i|$ is the number of images in the ith segment of the first-step segmentation results. Then, the mean of HSV histograms among M_i images is referred to the Segment-HSV feature of S_i, denoted by S-HSV.

E. Image data Segmentation Step

Two image processing modules are designed in the second data processing step. Briefly, Image Processing Module I (IPM-I) is designed to determine whether there are new activities within a long segment, while IMP-II handles the cases where a single SS segment or multiple continuous SS segments are combined to form a new segment.

Image Processing Module I: In this module, all images belong to a long active or inactive segment according to the motion sensor data. However, for some cases, it is possible that more than one activities occur in this segment. For instance, after walking outside for a long time, a subject enters a mall for shopping. Walking outside and shopping are considered two different activities, but they are not differentiable from the motion sensor data.

In order to detect activities hidden in motion features, we equally cut the segment S_i into sub segments from P_1 to P_k in every 5 minutes, where k is the number of sub segments. Image feature distance between adjacent subsegments is calculated. If $D\left(P_k,\ P_{k-1}\right) \geq T_{hist}$, the boundary between two subsegments is marked as a new boundary, where $D\left(P_k,\ P_{k-1}\right)$ is the distance between P_k and P_{k-1}, and T_{hist} is a threshold set experimentally as described in Section III.

Image Processing Module II: Our segmentation goal in this module is to help subject recall major events during the day, rather than short and insignificant events. When the duration of SS segments is less than threshold T_d, for both S_{i-1} and S_{i+1}, there are two conditions: (1) if one label is AS and the other label is IS, segment S_i will be merged with the AS segment; (2) If the labels of S_{i-1} and S_{i+1} are the same, the image feature distance between S_{i-1} and S_{i+1} is calculated. If $D\left(S_{i-1},\ S_{i+1}\right) \geq T_{hist}$, S_i will be merged with S_{i-1}, otherwise S_{i-1}, S_i and S_{i+1} will be combined into one single segment.

When the duration of continuous SS segments is equal to or greater than threshold T_d, We will determine which segments will be merged based on the similarity measure between the current and previous/next segments. If the current segment is determined to be different from any of the neighboring segments, it is denoted as one independent segment. Otherwise, it is combined with one neighboring segment. Specifically, we calculate image feature distance $D(S_{i-1}, S_i)$ and $D(S_{i+1}, S_i)$ according to the following criteria:

If $D(S_{i-1}, S_i) < D(S_{i+1}, S_i)$ and $D(S_{i-1}, S_i) < T_{hist}$
S_i is merged with S_{i-1}
Else If $D(S_{i+1}, S_i) < D(S_{i-1}, S_i)$ and $D(S_{i+1}, S_i) < T_{hist}$
S_i is merged with S_{i+1}
Else S_i is identified as an independent segment

Once all segments are processed, the final segmentation results are obtained and presented to the user for manual identification of the activity within each segment.

III. Experiments

A. *Datasets*

In our experiments, a total of twelve-day data were recorded from six human subjects, including weekdays and weekends. For each subject, two days of data were selected according to the completeness of the datasets. All subjects were not asked to modify their usual daily activities and the data were captured under free-living conditions. The duration of each selected dataset was longer than ten hours. Each sequence consisted of gyroscope data, accelerometer data, and more than 10,000 images.

All six subjects reviewed their own test data and marked activity boundaries manually as the ground truth. To evaluate the performance of the proposed method, the following measures were utilized:

$$\text{Recall} = \frac{N_c}{N_c + N_m} \qquad \text{Precision} = \frac{N_c}{N_c + N_f}$$

$$F_1 = \frac{2 \times \text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}} \qquad (1)$$

where N_c is the number of boundaries in the ground truth, N_m is the number of missing boundaries, and N_f is the number of false boundaries. F_1 is a performance measure widely utilized in pattern recognition and information retrieval fields. It is the harmonic mean of both recall and precision [15].

B. *Threshold determination*

Two thresholds, including T_m for the motion feature and T_{hist} for the image feature, are important parameters distinguishing boundaries of activities. These parameters were determined by the following empirical means. Since we have large amounts of data recorded, we utilized ten days of data to manually classify between active and inactive activities. Then, the accuracies under different threshold values were calculated using the manual classification result as the gold standard. The top curve in Fig.3 shows the accuracy in terms of S-STD computed from the gyroscope data versus the threshold value. It can be observed that T_m equal to 0.151 provides the best performance.

Similarly, in the image feature case, we determined threshold T_{hist} by manually segmenting the same ten days of data into activity sequences. Two sets of similarities were calculated. Firstly, each single segment was equally divided into two sub-segments, and the similarity between these two sub-segments was calculated to form the first set. Next, similarities between adjacent segments were calculated and grouped into the second set. The threshold which best separates these two sets was determined to be T_{hist}. The bottom curve in Fig.3 shows the accuracy under different threshold values. It can be observed that, when T_{hist} equals to 0.173, the accuracy is the highest (66.73%).

C. *Test on different motion sensor features*

In our experiments, the features computed from different types of motion sensors were compared in order to demonstrate which sensor or sensor combination provides the most information. Table I shows the F_1 values defined in (1) computed using different motion sensors. In this table, GYR and ACC denote, respectively, the motion data acquired by the gyroscope and accelerometer. DER denotes a different feature proposed in [4] which is a combination of derivatives along three axes of acceleration, smoothed by a Gaussian filter using a window function. From Table I, the F_1 value of S-STD of the gyroscope data provides the best results. Besides the S-STD of the gyroscope, DER of the accelerometer also provides good results. It is thus concluded that both the gyroscope and accelerometer data are suitable choices to characterize motion in the recorded daily activity.

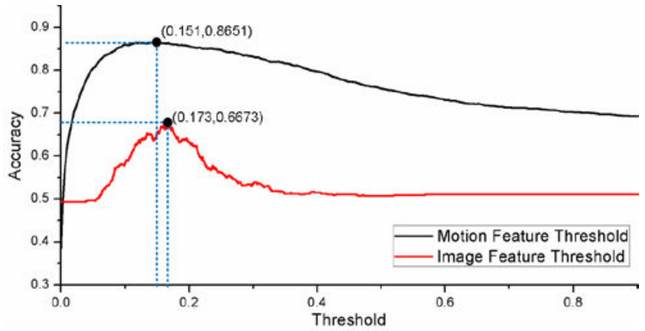


Figure 3. Accuracies based on motion and image features versus thresholds

TABLE I. PERFORMANCE ON DDIFFERENT MOTION FEATURES

	Recall	Precision	F_1
GYR+DER	0.63	0.76	0.69
ACC+(S-STD)	0.67	0.71	0.69
ACC+DER	0.76	0.67	0.72
GYR+(S-STD)	0.82	0.69	0.75

D. Test on Different window Sizes and Overlaps

The effects of different data sampling window sizes and their overlaps were evaluated in our experiments. The windows sizes tested were 16, 32, and 64 seconds, and the overlaps tested were 0%, 50% and 75%. Our results are shown in Table II. It can be observed that the 50% overlap and the 32 second window size performed the best among all choices. It can also be observed that the cases of a small window with any overlap and any window without overlap both yield poor performance. It may be explained that the noise effect may be responsible for the performance reduction in these cases. As seen from Table II, for the 64-second window case, the result is slightly inferior to that of the 32-second window. This may be caused by the loss of some detail information due to the excessive large window size.

E. Evaluation between single sensor and proposed method

Finally, we conducted an experiment to compare our method with the methods using a single type of sensor data (gyroscope data or the image data). Fig.4 presents F_1 curves of the three methods with respect to twelve days of data (horizontal axis). It can be seen that our method provides the best result. Although our method examines both types of data, its computational cost is still low because the image processing procedure is performed only using a small portion of the image data in the neighborhood of candidate activity boundaries.

IV. CONCLUSION

A two-step method for multi-sensor data segmentation has been proposed. In this method, the S-STD feature of the gyroscope is first extracted to generate a set of candidate boundaries. Then, the segment-HSV feature is utilized to fine-tune the result in the neighborhood of candidate boundaries to enhance performance. Using the proposed method, the multi-sensor daily life data recorded by a wearable device can be automatically segmented into a relatively small set of activities, which greatly reduces the task of activity recognition which is currently conducted manually. Automatic daily event recognition algorithms are currently being studied by our and other research groups.

TABLE II. PERFORMANCE ON DIFFERENT WINDOW SIZES AND OVERLAPS

Length(s)	Overlap	Recall	Precision	F_1
16	75%	0.73	0.57	0.64
16	50%	0.75	0.62	0.68
64	0%	0.72	0.65	0.68
16	0%	0.77	0.63	0.69
32	0%	0.76	0.69	0.72
64	50%	0.79	0.68	0.73
64	75%	0.8	0.68	0.74
32	75%	0.79	0.69	0.74
32	50%	0.82	0.69	0.75

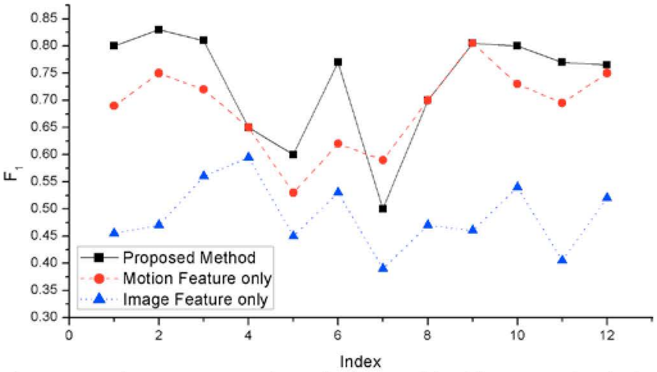


Figure 4. Performance comparison of using combined features and a single feaure

REFERENCES

[1] I.C. Gyllensten and A.G. Bonomi, "Identifying Types of Physical Activity with a Single Accelerometer: Evaluating Laboratory-trained Algorithms in Daily Life," *IEEE Trans. on Biomedical Engineering*, vol.58, no.9, pp.2656-2663, 2011.

[2] T. Zhang, W. Tang and E.S. Sazonov, "Classification of posture and activities by using decision trees," in *IEEE International Conference on Engineering in Medicine and Biology Society*, pp.4353-4356, 2012.

[3] K.Tong and M.H. Granat, "A practical gait analysis system using gyroscopes," *Medical Engineering & Physics*, vol. 21, no. 2, pp.87-94, 1999.

[4] A.R. Doherty, A.F. Smeaton, K. Lee and D. P. Ellis. "Multimodal segmentation of lifelog data," in RIAO 2007 - Large-Scale Semantic Access to Content (Text, Image, Video and Sound) 2007.

[5] M. Sun, J. Fernstorm, W. Jia, S.A. Hackworth, N. Yao, Y. Li, C. Li, M.H. Fernstrom and R. J. Sclabassi, "A wearable electronic system for objective dietary assessment" *Journal of the American Dietetic Association*, vol. 110, pp. 45-47, 2010.

[6] eButton description. Available: http://www.lcn.pitt.edu/ebutton/

[7] P. Browne, A. F. Smeaton, N. Murphy, N. O'Connor, S. Marlow and C. Berrut, "Evaluation and combining digital video shot boundary detection algorithms," in *Proc. Irish Machine Vision and Image Processing Conference*, pp. 108–114, 2000.

[8] C. Grana and R. Cucchiara, "Linear transition detection as a unified shot detection approach", *IEEE Trans. on Circuits System for Video Technol.*, vol. 17, no. 4, pp. 483–489, 2007.

[9] Z. Rasheed and M. Shah, "Scene detection in hollywood movies and TV shows," in *Proc. IEEE International Conference on Computer Vision and Pattern Recognition*, vol.2, pp.343-348, 2003.

[10] Z. Cernekova, I. Pitas and C. Nikou, "Information theory-based shot cut/fade detection and video summarization," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 16, no. 1, pp.82-91, 2006.

[11] W. Tavanapong and J. Zhou, "Shot clustering techniques for story browsing," *IEEE Trans. on Multimedia*, vol.6, no.4, pp. 517–526, 2004.

[12] H. Feng, W. Fang, S. Liu and Y. Fang, "A new general framework for shot boundary detection based on SVM," in *Proc. International Conference on Neural Networks and Brain*, vol.2, pp.1112-1117, 2005.

[13] A.M. Khan, Y. Lee, S.Y. Lee and T. Kim, "A triaxial accelerometer-based physical-activity recognition via augmented-signal features and a hierarchical recognizer," *IEEE Trans. on Information Technology in Biomedicine*, vol.14, no.5, pp. 1166-1172, 2010.

[14] S. Chang, T. Sikora and A. Purl, "Overview of the MPEG-7 standard," *IEEE Trans. on Circuits and Systems for Video Technology*, vol.11, no.6, pp.688-695, 2001.

[15] D. M. W. Powers, "Evaluation: from precision, recall and F-factor to ROC, informedness, markedness & correlation," *Journal of Machine Learning Technologies,* vol.2, no.1 pp. 37–63, 2011