Automated Habit Detection System: A Feasibility Study

Hiroki Misawa, Takashi Obara, and Hitoshi Iyatomi(✉)

Applied Informatics, Graduate School of Science and Engineering,
Hosei University, Tokyo, Japan
iyatomi@hosei.ac.jp

Abstract. In this paper, we propose an automated habit detection system. We define a "habit" in this study as some motion that is significantly different from our common behaviors. The behaviors of two subjects during conversation are tracked by the Kinect sensor and their skeletal and facial conformations are detected. The proposed system detects the motions considered as habits by analyzing them using a principal component analysis (PCA) and wavelet multi-resolution analysis (MRA). In our experiments, we prepare a total of 108 movies containing 5 min of conversation. Of these, 100 movies are used to build the average motion model (AMM), and the remainder are used for the evaluation. The accuracy of habit detection in the proposed system is shown to have a precision of 84.0 % and a recall of 81.8 %.

1 Introduction

Although we primarily communicate in words, it is well known that non-verbal communication such as expressions, gestures, and unconscious motions have a significant influence on our communication [1]. Habits are behaviors that we often perform unconsciously or have little awareness. Some habits might make people displeased and, in some cases, could be the cause of a loss of opportunity. Thus, we consider the objective recognition of our habits to be meaningful not only for better communication, but also for a wide range of general purposes.

Many studies on motion analysis have considered a wide range of objectives. The methodologies can be divided into two categories from the perspective of the usage of sensory devices: (1) subjects wear sensory devices and their motions are estimated based on obtained signals, or (2) subjects wear no special devices and their motions are directly estimated from video recordings with image processing techniques. In the former, acceleration sensors are commonly used [2] because of their excellent practical applicability in detecting gradient, motion, and fluctuation. These sensors provide meaningful information, although they are sometimes unavailable because of limitations in terms of cost, weight, and geometry. In the latter category, commercially available video cameras have been widely used. Bobick and Davis [3] identified the behavior of subjects by extracting the transformation areas of the silhouette from movies and generating the binary

G. Bebis et al. (Eds.): ISVC 2015, Part II, LNCS 9475, pp. 16–23, 2015.
DOI: 10.1007/978-3-319-27863-6_2

motion energy image and motion history image. Schuldt et al. [4] also identified subjects' motion by means of a bag-of-features consisting of a histogram of local features and a support vector machine (SVM) as a classifier. Infrared cameras have also been used for motion analysis, either alone or in conjunction with visible light cameras, because of their tolerance for variations in lighting conditions. In each of these cases, some depth estimation of the target is necessary when there is a need for 3D analysis. In many cases, depth estimation is achieved by using a stereopsis system with multiple cameras. If highly accurate analysis is required, marker detection is commonly used, although this method needs dedicated equipment and/or facilities.

The Kinect sensor was released in 2010 by Microsoft Corp. as a peripheral device for their Xbox gaming platform. As the Kinect is relatively cheap, but can track 3D motion with a considerable measure of credibility, it has been used in many studies [5–7]. Xia et al. [5] recognized human behavior with their HOJ3D (Histograms of 3D joint locations) method. This method generates 3D histograms of human posture that are analyzed by a linear discriminant analysis and a vector quantization. Evangelidis et al. [6] performed motion recognition with Fisher vectors based on the location of articulations obtained from 3D skeletal information and the SVM classifier. Miranda et al. [7] proposed a gesture recognition system. Their method detected characteristic motion from the observed 3D skeletal information with the SVM, and classified the detected motion using a trained decision tree. However, to the best of our knowledge, no systematic research on automated habit detection or habit analysis has been conducted. We believe this is because the variety of habits is quite broad, making it difficult to apply conventional methodologies.

In this study, we propose an automated habit detection system that utilizes the Kinect sensor. Considering the further applicability of this system, we analyze the behavior of the subjects during conversation.

2 Habit Detection System

First, we define a habit as some motion that is significantly different from our common behaviors. The proposed habit detection system tracks the behavior of two subjects in conversation with the Kinect sensors, and detects distinctive behavior as a habit. The schematics of the proposed habit detection system are shown in Fig. 1. The proposed system has two operation phases: a training phase and an evaluation phase. In the training phase, we record a large number of conversations with the Kinect sensor and form the average motion model (AMM) for each body part as a reference. In the evaluation phase, the behavior of the subject is compared with the AMM, and significant differences are identified as habit. The details will be explained in later sections. The proposed system calculates the velocity of the subjects' body parts. The time-series of the velocity of each body part is analyzed using a wavelet multi-resolution analysis (MRA).

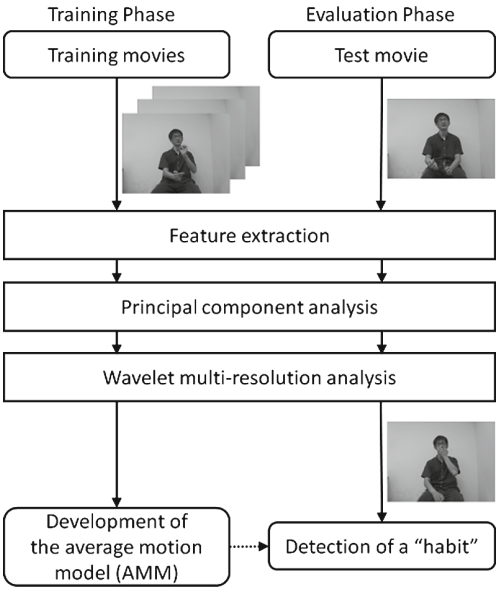

Fig. 1. Schematics of the proposed habit detection system.

2.1 Recording Environment and Material

In this study, we recorded video clips as follows: (1) Two subjects sit face-to-face, (2) the Kinect sensor is set up in front of each subject to track his/her upper body, and (3) their conversation is recorded with the Kinect sensor for at least 300 s. The recording environment is shown in Fig. 2. The Kinects are approximately 1 m from their respective subjects, and the distance between the two subjects is approximately 2.5 m. We prepared several topics for conversation (e.g., school life, hobbies, friends), and the subjects selected one of these topics prior to the recording.

In this experiment, we recorded three movies of each of 36 male subjects (24.3 ± 1.3 years old), i.e., a total of 108 movies, and selected an arbitrary 300 s from each movie for processing. We used the Kinect Studio (SDK 1.5) to record the conversations at 30 fps and detect the 3D motion information. The RGB and depth sensor resolutions of the Kinect were 640×480 and 320×240 pixels, respectively.

2.2 Detection of Tracking Points

In this study, we used skeletal information of the upper body and the facial components of the subjects. For the former, we detected a total of 10 joints from the upper body while the subjects were seated (Fig. 3). For the latter, we used five points (forehead, left eye, right eye, nose, and mouth; see Fig. 4) out of 121 detected facial feature points. Accordingly, a total of 15 three-dimensional

Fig. 2. Experimental environment.

feature points (i.e., 45 features) were extracted from each image frame for each subject. We calculated the velocity of each point by investigating the difference in point locations between successive frames. Accordingly, the 3D velocity vector for each body part p ($p = 1, 2, \cdots, 15$) at time t is expressed as:

$$\mathbf{v^P(t)} = [v_x^p(t), v_y^p(t), v_z^p(t)]^T. \tag{1}$$

Because the motion of the body has physical and geometrical limitations, we conducted a principal component analysis (PCA) for each velocity vector. Table 1 summarizes the contribution of the first primary component in each body part. According to these results, we can confirm that the first primary component makes a significant contribution to many body parts. Therefore, we decided to approximate the obtained 3D velocity $\mathbf{v^P(t)}$ by the one-dimensional velocity value $v_{1st}^p(t)$ in the direction of the first eigenvector. In the experiment, we analyzed these 15-dimensional time-series data

$$\mathbf{V(t)} = [v_{1st}^1(t), v_{1st}^2(t), \cdots, v_{1st}^{15}(t)]^T \tag{2}$$

by means of MRA with Daubechies' wavelet (N = 2). According to the results of the preliminary experiment, we focused on wavelet coefficients with a frequency of 1.25 Hz for each body part.

2.3 Definition of "habit" and Its Detection

In the training phase, we formed the AMM of each subject's body parts by averaging the wavelet coefficients of the training dataset (i.e., 100 movies). In the evaluation phase, wavelet coefficients were calculated for each subject's body motions. If the difference between the evaluation target and the AMM was greater than twice the standard deviation (SD) of the AMM, the system considered this motion to be uncommon, and therefore identified it as a habit.

To conduct a quantitative evaluation of the proposed system, the gold standard is required. As there is no objective definition of a habit in terms of physical motions, the authors manually selected several motions considered as habits from

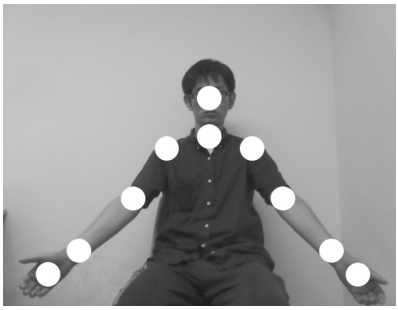

Fig. 3. Ten tracking points in upper body.

Fig. 4. Five tracking points on the face.

Table 1. Contribution ratio of the PCA.

Body part	Contribution ratio (%)	Body part	Contribution ratio (%)
Head	82.6	ShoulderCenter	79.6
ShoulderLeft	81.1	ShoulderRight	79.7
ElbowLeft	64.7	ElbowRight	65.3
WristLeft	45.4	WristRight	45.9
HandLeft	47.6	HandRight	50.6
Forehead	88.1	Nose	88.2
EyeLeft	89.1	EyeRight	89.4
Mouth	87.9		

the evaluation dataset, and determined them as the gold standard. We used the precision and recall as performance criteria. These were calculated as follows:

$$\text{precision} = \frac{\#\text{ of correctly detected habit}}{\#\text{ of detected habit}}. \tag{3}$$

$$\text{recall} = \frac{\#\text{ of correctly detected habit}}{\#\text{ of habit (gold standard)}}. \tag{4}$$

If a habit was detected by the system within 2 s of the gold standard, we considered the detection to be appropriate.

3 Results

In this experiment, we used a total of 100 movies to build the AMM, with the reminder used for evaluation. We determined 77 motions as habits for the gold standard. We illustrate our results with an example. Figure 5 is a sample diagram summarizing the habit detection results of our system and the gold standard. Figures 6 and 7 are examples of true positives, Fig. 8 is a false positive, and Fig. 9 is a false negative. In our system, the body part considered to have a habitual

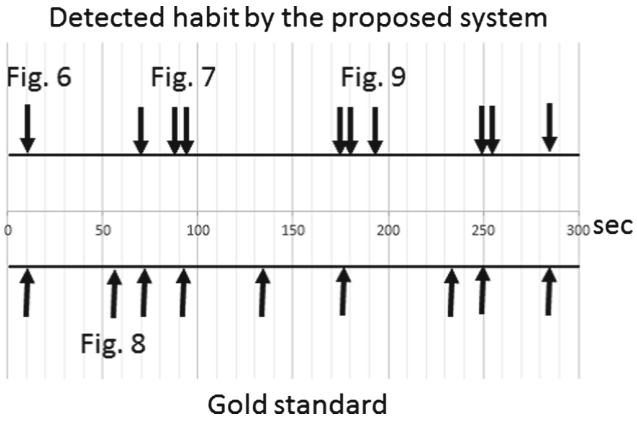

Fig. 5. Comparison of habit detection.

Fig. 6. Example of correct detection 1 (6.1 SD from the AMM).

Fig. 7. Example of correct detection 2 (2.4 SD from the AMM).

motion is highlighted by a circle on the detected body part. In Figs. 6 and 7, the subject scratches his nose with his right hand (at 10 s, 6.1 SD from the AMM), and leans backward (at 90 s, 2.4 SD). These motions were identified as being habits. Recall that the system detects those motions that differ from the AMM by more than 2 SD to be habits. In Fig. 8, we show an example in which the system falsely detected a commonly seen motion as a habit (at 60 s, 2.1 SD). On the contrary, Fig. 9 shows an example in which the system did not detect the habit of falling forward (at 190 s, 0.6 SD). The confusion matrix of our results is shown in Table 2. In summary, the habit detection performance of our system achieved a precision of 84.0 % and a recall of 81.8 %.

4 Discussion

In our experiment, the proposed system achieved good habit detection performance. It can be considered that the proposed methodology has the potential

Fig. 8. Example of false positive (2.1 SD from the AMM).

Fig. 9. Example of false negative (0.6 SD from the AMM).

Table 2. Confusion matrix in detection of habits.

	Gold standard		
Proposed system		habit	non-habit
	habit	63	12
	non-habit	14	

to detect human habits. However, we recognize that several issues need to be addressed to make our system practical. In this experiment, we detected only uncommon motions as habits, focused only on the speed of the motion, and employed a subjective gold standard. We will investigate these issues and develop an improved system in the near future.

5 Conclusion

In this study, we proposed a prototype automated habit detection system to objectively recognize unconscious habits. We used a total of 108 video clips of subjects in conversation, and achieved habit detection accuracy with a precision of 84.0 % and a recall of 81.8 %. Future work will focus on improving the accuracy and reliability of our methodology.

References

1. Merabian, A.: Communication without words. Psychol. Today **2**, 53–55 (1968)
2. Sagawa, K., Abo, S., Tsukamoto, T., Kondo, I.: Forearm trajectory measurement during pitching motion using an elbow-mounted sensor. J. Adv. Mech. Des. Syst. Manuf. **3**, 299–311 (2009)
3. Bobick, A.F., Davis, J.W.: The recognition of human movement using temporal templates. IEEE Trans. Pattern Anal. Mach. Intell. **23**, 257–267 (2001)

4. Schuldt, C., Laptev, I., Caputo, B.: Recognizing human actions: a local SVM approach. In: Proceedings of the 17th International Conference on Pattern Recognition (ICPR), vol. 3, pp. 32–36 (2004)
5. Xia, L., Chen, C.C., Aggarwal, J.K.: View invariant human action recognition using histograms of 3D joints. In: 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 20–27 (2012)
6. Evangelidis, G., Singh, G., Horaud, R.: Skeletal quads: human action recognition using joint quadruples. In: 2014 22nd International Conference on Pattern Recognition (ICPR), pp. 4513–4518 (2014)
7. Miranda, L., Vieira, T., Martinez, D., Lewiner, T., Vieira, A.W., Campos, M.F.M.: Online gesture recognition from pose kernel learning and decision forests. Pattern Recogn. Lett. **39**, 65–73 (2014)