Dynamic gesture classification using skeleton model on RGB-D data

Y. Tamura\textsuperscript{1}, T. Umetani\textsuperscript{1}, N. Kashima\textsuperscript{2} and H. Nakamura\textsuperscript{3}

\textsuperscript{1} Department of Intelligence and Informatics, Konan University, Okamoto, Higashinada, Kobe, 658-8501, Japan
\textsuperscript{2} Department of Energy Engineering and Science, Nagoya University, Frocho, Chikusa, Nagoya, 464-8603, Japan
\textsuperscript{3} Department of Helical Plasma Research, National Institute for Fusion Science, Oroshi, Toki, 509-5292, Japan

E-mail: tamura@konan-u.ac.jp

Abstract. This study aims to subjectively detect and classify similar gestures using a red-green-blue-depth camera. Human gesture recognition is one of the crucial components for realizing natural user interfaces (NUIs) using computers and machines. The quality of the NUI highly depends on the robustness of the achieved gesture recognition. We, therefore, propose a gesture classification method using singular spectrum transformation. Using this method, we can robustly classify gestures and behavior.

1. Introduction
Dynamic gesture classification is an important and growing area in fields such as human interface, robotics, and virtual reality. One obstacle encountered in automatically detecting gestures is that the raw data include noise; second, there are some representations to show one’s decision by gestures. To overcome these problems, we propose a gesture classification method using singular spectrum transformation (SST) \cite{1}.

This method has been applied to different fields, including earth sciences, physics, and financial research. In human interface research, this method has been applied to the segmentation of human motions \cite{2}. We propose a method to classify gestures using similar factors involving the results of SST.

2. System configuration
Our system consists of a Microsoft Kinect sensor and a PC. The sensor has a red-green-blue and a depth sensor; such sensors detect the position of not only an object but also human joints. The Kinect sensor can detect three degree-of-freedom (DOF) positions of 20 joints approximately 30 times per second. We used only six joints for gesture detection: elbows, shoulders, and knees. This translates to 18 DOF position data being sent from the Kinect sensor to the PC. The key reason for using only six joints is to reduce the required processing time; furthermore, most joint positions (e.g., the spine) are not related to gestures.
3. Method

Figure 1 shows the procedure governing our proposed system.

![Diagram of the procedure to detect gestures and behavior]

Figure 1. Procedure to detect gestures and behavior

In this study, we use SST for gesture classification. SST is a nonparametric decomposition method suitable for the decomposition of unsteady datasets. Consider one-dimensional time series dataset $x(1), x(2), x(3), \ldots, x(t)$. A vector $j$ having $w$ components is defined as

$$j(t-1) = [x(t-w), x(t-w+1), \ldots, x(t-1)]^T.$$  \hspace{1cm} (1)

The Hankel matrix (i.e., trajectory matrix) $H(t)$ having $n$ columns is described as

$$H(t) = [j(t-n), j(t-n+1), \ldots, j(t-1)].$$  \hspace{1cm} (2)

This matrix can be decomposed using singular value decomposition as

$$H(t) = U(t)W(t)V(t)^T.$$  \hspace{1cm} (3)

We apply this method to the time series of joint vector data in both the past dataset and the future dataset. The past dataset is composed of data from time $(t-n-w-1)$ to $(t-1)$, whereas the future dataset is composed of data from $(t+s)$ to $(t+s+w+n)$, where $s$ is the time shift and $w$ and $n$ are parameters mentioned in eq. 1 and 2, respectively. Comparing results of both datasets, we can extract the motion change point.

For detecting motion change, we introduce "the degree of change" factor. A hyperplane $A_p(t)$ is defined using $n$ largest left-singular vectors ($u_{p1}, u_{p2}, \ldots, u_{pn}$) from past matrix $U_p(t)$ by eq. (3). The largest left-singular vector $u_f(t)$ is calculated from future matrix $U_f(t)$. If the tendency of the past and the future dataset is similar, the inner product between vector $u_f(t)$ and the vector, which is the projected $u_f(t)$ to hyperplane $A_p(t)$, is approximately 0. In the opposite case, the factor becomes 1. We define the degree of change by the following equations:

$$S(t) = 1 - u_f(t)P(t)$$  \hspace{1cm} (4)

$$P(t) = A_p(t)A_p(t)^T,$$  \hspace{1cm} (5)

where all components of matrix $A_p$ are orthonormal bases.

Next, we estimate similarity using the Hillinger distance between a control gesture and the other gesture as follows:

$$d(i,j)^2 = \int (\sqrt{S_i(t)} - \sqrt{S_j(t)})^2 dt.$$  \hspace{1cm} (6)
| Case | Gesture Description               |
|------|----------------------------------|
| 1    | Right-hand gesture - Stand still |
| 2    | Right-hand gesture - Walk        |
| 3    | Left-hand gesture - Stand still  |
| 4    | Right-hand up - Stand still      |
| 5    | Right-hand gesture - Stand still |

**Table 1.** Gestures for estimation

**Figure 2.** Gestures using the right hand and arm corresponding to cases 1 and 5 of Table 1

### 4. Results

Table 1 shows gestures for estimation of the proposed method. The user stands in front of the Kinect sensor approximately 3 m away. Except case 2, the user gestures while standing still; however, in case 2, the user gestures by walking. In cases 1 and 5, the user gestures to indicate the right direction; however, these gestures differ slightly, as shown in Fig. 2. The parameters used in eq. 1 and 2 are $w = 30$, $n = 30$, and $s = -15$. Parameter $w$ needs to be set sufficiently long to include all gesture signals; therefore, $w$ is set to 30, indicating that the time range of the matrix is approximately 1 s. In this experiment, we used only $u_{p1}, u_{p2}$ to obtain hyperplane $A_p$. Figure 3 shows results of case 1. The left-hand image shows raw data of the elbow joint direction, and the right-hand image shows the SST results. In this case, the user gestured three times, and results show the gestures successfully extracted from the raw data.

**Figure 3.** Successful extraction of meaningful gestures: on the left, the $x$ component of raw data of the right elbow to right wrist; on the right, results of extraction using SST

Figure 4 shows results of case 2. In this case, the user gestured four times while walking. In general, it is difficult to extract distinct gesture signals and walking signals; however, our system successfully extracts the gesture signals. Moreover, as shown in the left-hand column of Fig. 4, if this method is applied to the left-hand movement (i.e., no gesture), no signal is detected. After extracting the gesture area, the next step is to detect what the extracted gesture denotes (i.e., classify the gesture). From eq. 6, we determine the gesture type. Figure 5 shows the classification results. The distance is calculated between the right-hand gesture (i.e., case 1 is the control gesture) and the other gestures. Results indicate that our method can successfully determine the joint, which has moved, and the gesture by the distance factor.

### 5. Conclusion

We applied SST to extract meaningful gestures and behavior from raw data obtained via a Microsoft Kinect system. The raw data are generally noisy, and therefore, it is difficult to extract useful data. SST removes continuous noisy data and extracts objective data with only
a few specified parameters. Furthermore, we successfully classified the meaning of the gesture using the Hillinger distance.

Acknowledgement
This study is supported by KAKENHI (22500114).

References
[1] J. B. Elsner and A. A. Tsonis 1996 Plenum Press
[2] H. Nakanishi, S. Kanata, H. Hattori, T. Sawaragi and Y. Horiguchi 2011 Journal of Advanced Computational Intelligence and Intelligent Informatics 15 8

Figure 4. Results of extraction in case 2 (gesturing with walking) in which the upper half shows raw data, and the lower half shows SST results: the left-hand and middle columns present the x and y vector components, respectively, from the right elbow to the right wrist; the right-hand column presents the y vector component of the left hand.

Figure 5. Classification results using the Hillinger distance: the upper-left graph shows the sum of all distances of all components (i.e., six joints); the upper-right graph shows the knee vector; the lower-left graph shows the right hand vector; and the lower-right graph shows the right shoulder vector.