Human Gesture Recognition using Keyframes on Local Joint Motion Trajectories

Rafet Durgut
Computer Engineering Department
Karabuk University
Karabuk, Turkey

Oğuz FINDIK
Computer Engineering Department
Karabuk University
Karabuk, Turkey

***Abstract*—Human Action Recognition (HAR) systems are systems that recognize and classify the actions that users perform against the sensor or camera. In most HAR systems, an input test data is compared with the reference data in the database using various methods. Classification process is performed according to the result obtained. The size of the test or reference data directly affects the operation speed of the system. Reduced data size allows a significant performance increase in system operation speed. In this study, action recognition method is proposed by using skeletal joint information obtained by Microsoft Kinect sensor. Splitting keyframes are obtained from the skeletal joint information. The keyframes are observed as a distinguishing feature. Therefore, these keyframes are used for the classification process. Keeping the keyframes instead of keeping the position or angle information of action in the reference database can benefit from memory and working time. The weight value of each keyframes is calculated in the method. The problem of temporal differences that occur when comparing test and reference action is solved by Dynamic Time Warping (DTW). The k-nearest neighbor's algorithm is used for classification according to the obtained results from DTW. The sample has been tested in a data set so that the success of the method can be tested. As a result, 100% correct classification was achieved. It is also suitable for working at real time systems. Breakpoints can also be used to provide feedback to the user as a result of the classification process. The magnitude and direction of the keyframes, the change in the trajectory of joint, the position and the time of its existence also give information about the time errors.**

Keywords—Human gesture recognition; dynamic time warping; local joint motion trajectory; Human action recognition; microsoft kinect

I. INTRODUCTION

Human Computer Interaction (HCI) is one of the areas that have been working hard in recent years [9]. HCI is directly interacting with human actions. Functions of monitoring, controlling and analyzing using human movements can be realized. Elderly and child individuals can be prevented from falling into dangerous situations by being kept under constant surveillance [11]. Controls of machines or robots can be provided by human actions. Analysis can be made by following the development process of the athlete or orthopedic patients. This and many other applications can be cited as examples of HCI issues [24]. Due to the increased sensitivity of the hardware that can be used in the HCI field, the accuracy and accuracy of the developed applications is also increasing. The quality of human life can be increased directly through the applications made on this area. As the main motivation of researchers in HCI field, positive contributions to human life can be shown.

Human Activity Recognition can be shown as a subdivision of the HCI field. Computer side developments provide positive contributions directly to the work done in this area. HAR is a method of recognizing Human Actions using various algorithms in the computer environment via cameras and sensors. As the accuracy and performance of the devices and equipment used in the process of recognizing the action increase, the availability and diversity of the work done increases. The development of RGB-D cameras in recent years and the increased sensitivity of these cameras have led to the use of many important field recognition actions [2]. New researchers have been working on this field, especially with interesting studies in health, safety, smart home systems, surveillance and control areas [3]. The most important advantages of HAR systems are that they can restore the motor skills of the person, make the physical therapy and rehabilitation exercises more feasible, and provide a fun environment [4].

During the Human Action recognition process, people perform their performance in front of the camera or sensor. This performance information is transferred to the computer environment through the hardware. Thanks to the sensors on the hardware, a 3D depth image can be obtained with RGB image. This data, called RGB-D, has become popular in the field of pattern recognition. In this way, the position information of the objects or persons on the real world can be reached. The important thing is that these 3D data can be obtained with low cost hardware.

With the widespread use of low-cost depth sensors, the number of human action recognition efforts using depth maps has also begun to increase [5]. Detection and classification of dance figures using the Kinect sensor can be performed in real time [6]. In drumkit simulator application, recognition and classification of drumkit actions can be realized in real time using kinect sensor [7].

Skeletal-joint based human gesture or action recognition applications use depth maps. Thanks to these depth maps obtained from depth cameras, skeletal-joint representation is used. Coordinates of joint points are used directly or indirectly in this field work. As the simplest and basic feature, the distances between the previous frame and the next frame can be used with reference to a coordinate system of the joints [8].

Changes in joints made in 3D space can also be used as a distinguishing feature [9]. Action recognition can be performed by using joint-joint angles joint-plane angles[10]. Various features are used as input parameters in human action recognition systems. The purpose of using different features is to get higher performance with higher accuracy. Using these features, many methods have been developed in the field of human action recognition. (HMM), Dynamic Time Warping (DTW), Support Vector Machine (SVM), Neural Networks (NN) [14], Logistic Regression (LR) [6], Adaptive Bayesian Models (ABM) [15], k-nearest neighbor (knn) [23] are the most commonly used methods in this area. These methods use human-joint representation as input parameters in human action recognition applications.

As the dimension of human-joint representation used as input parameter in human action recognition systems increases, the performance decreases. Joint-joint-based matching algorithms are not suitable for real-time applications due to this reason. Dimension reduction methods such as Principal Component Analysis (PCA) [16] and Linear Discriminant Analysis (LDA) [17] can be used to remove this problem altogether. Another solution is to use features with less size instead of using all the joint information from the position information in the human body skeleton representation. Instead of the trajectory of all body joints, action recognition can be performed using the trajectory of hand joint and the shoulder position [7]. The use of only significant joint trajectory rather than using all the joint information, can provide a significant increase in performance. Action recognition can also be performed by normalizing the angles of local joints and applying them as input to modified spherical harmonics (MSHs). The obtained information creates the angular skeleton system of action in the light. Then the joints in this system can be represented by the spherical coordinate system unit sphere function [18]. An action recognition system can also be realized by using the features of the most important and most informative joints in the series of joints that constitute an action [19].

In this study, local keyframes in the trajectory of the skeletal-joint positions were determined and used as input parameters in the comparison process. The local keyframes that make up the reference action and the local keyframes that make up the test action are compared using the DTW method. The obtained comparison results are applied as an input to K-NN classification algorithm and classified. The keyframes that make up an action have a distinguishing feature for comparison. By means of these keyframes, the errors between the reference action and the tested action can be determined in terms of time and amplitude and feedback can be provided to the user. It is suitable for real time operation due to its low memory requirement and high operating speed.

II. Proposed Study

In the action recognition process, the user performs his / her performance against depth cameras such as Kinect v2 sensor. Kinect sensor determines 20 joint position according to the user's skeletal joint representation. Each joint transfers its position over 3D space (x, y, z). According to the user's movements, these 20 joint position are kept in computer memory. The transfer is repeated 30 times per second [20]. The input parameters for action recognition applications are the series of joint position obtained as a performance result. Properties are extracted by performing various operations on the body joints sequences. The comparison is carried out by holding these properties in the test and reference motion sets. The test movement is then placed in the appropriate class. The block diagram of the system used to perform the human action recognition in this study is shown in Fig 1.

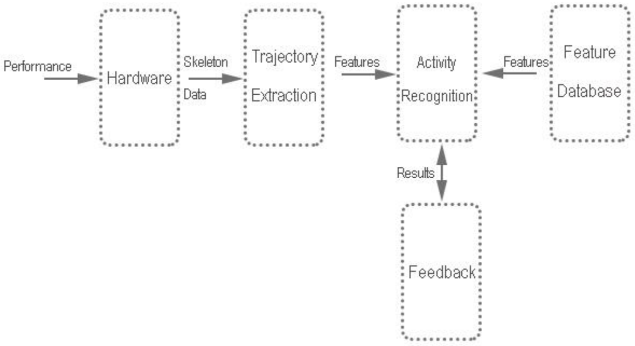

Fig. 1. General structure of the system

In a system with a total of 5 blocks, the flow is performed in sequential order. The performance information performed by the user as an input to the system is applied. This performance information is presented as feedback in the form of action information on the user through the devices and software.

A. Hardware

The part where the camera or sensor (e.g. Kinect) is used. In this section, user actions are performed in front of the sensor. The sensor creates a 3D depth map using the 2D images it receives. With these 3D depth maps, the position information of the joints of the person performing the action is obtained. The position information obtained is formed from the joint points shown in figure 2.

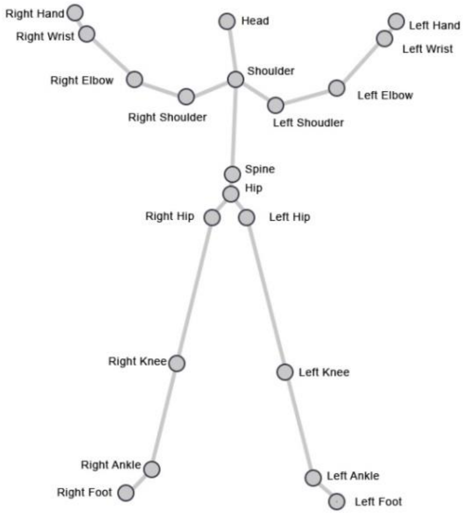

Fig. 2. Kinect Skeletal Joint Representation

This location information is transferred to the next block diagram in the form of skeleton data.

The positional information of a sample joint coordinate value (Right Hand Z coordinate) is shown in Fig 3. The action consists of a total of 48 frames.

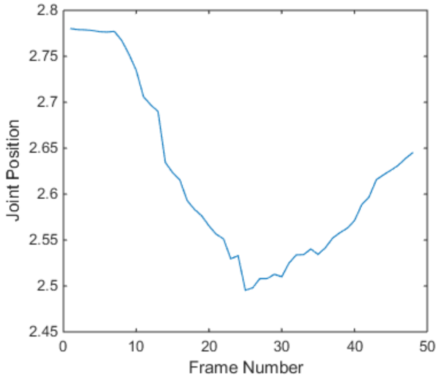

Fig. 3. Joint Coordinate Curve

The position $P_{i,t}$ of any i at any time t is kept in the series $P^{x}_{i,t}$ represents the position of the joint in the x coordinate system.

B. *Local Breakpoint Extract*

In order for the action recognition process in the system to work, it is necessary to detect the breakpoint first. The input time series used in this detection process is F^{p}. For the 20 joint points in the F^{p} time series, there are 60 pieces of coordinate information on the 3D coordinate plane.

$$F^{p}(t_0 : t_n) = < P^{x}_{i,t}, P^{y}_{i,t}, P^{z}_{i,t} > \qquad (1)$$

For each coordinate (*x*, *y*, *z*), coordinate information of each joint is applied as input to the system via hardware and API. The time series F^{d} holds the change value in each frame of the input series F^{p}.

$$F^{d} = F^{p}(t_{i+1}) - F^{p}(t_i) \qquad (2)$$

F^{bp} time series holds the local minimum and maximum values for F^{d} time series. a is the starting minimum for the local minimum point, *b* is the ending point. In that case;

$$F^{bp}_{i,t} = \sum_{j=a}^{b} F^{d}_{i,j} \qquad (3)$$

$$F^{bpp}_{i,t} = \begin{cases} t, & F^{bp}_{i,t} > F^{T} \\ 0, & otherwise \end{cases} \qquad (4)$$

The obtained $F^{bp}_{i,t}$ and $F^{bpp}_{i,t}$ features F^{b}_{i} are the keyframes. $F^{bp}_{i,t} i$. the size of the fracture point of the joint is kept. If $F^{bpp}_{i,t}$, the frame number of the fracture point of the *i-th* joint is retained. Thanks to these features, we have information about each action. $F^{bpp}_{i,t}$ gives information about the time of the change in the relevant joint, while $F^{bp}_{i,t}$ gives information about the magnitude of these changes. If the magnitude of this change is smaller than the threshold value F^{T}, no breakpoint is added. The threshold value can be selected according to the feature of the system to be applied manually, or it can take different values according to the given data set such as average, average of positives, average of negatives. In this work the threshold value is set manually.

Many features can be obtained from skeletal information. Speed of movement, direction of movement, acceleration of joint, etc. We can obtain information from this skeletal knowledge. At the same time, when we examine all of this skeleton information, certain moments (frame intervals) can make more sense in terms of motion. In this study, these points are called break points. Figure 4 shows the fracture points of the joint points given in Figure 3. The break points feature allows you to determine which class the motion belongs to, without examining the entire time series that makes up the movement.

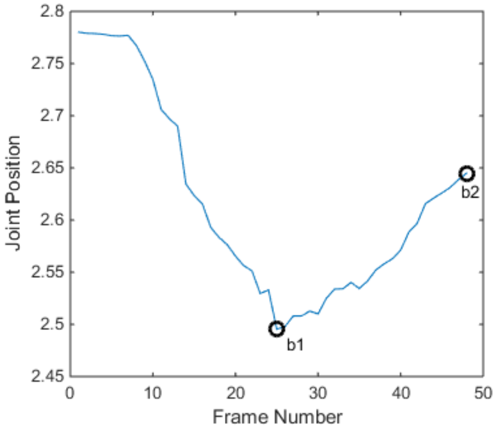

Fig. 4. Detection of break points

C. *Feature Database*

A reference training set is needed for the recognition or classification of movements. The performance characteristics are determined by comparing the characteristics of each reference movement in the training set. Generally, in order to be able to represent a movement in the database, recording may include position information of all the joints, as well as the only necessary features. In this way, the size of the database can be reduced. The feature database in this study contains only the required breakpoint properties instead of all the joint points.

D. *Gesture Recognition*

The breakpoint property information obtained from the performance performed by the user is compared with the breakpoint property information of the previously recorded actions in the database. Error rates and similarity ratios are calculated after comparison. As a result of this information, theaction class to which the performance belongs to is determined.

The properties obtained from the time series that are obtained as a result of the performance performed by the user are compared with the property values calculated by the only one-time algorithm in the database. Similarity and error rates of actions arise as a result of this comparison process. In Figure 5, when the same action is performed by different users, the change graph and fracture points at the same joint points are given.

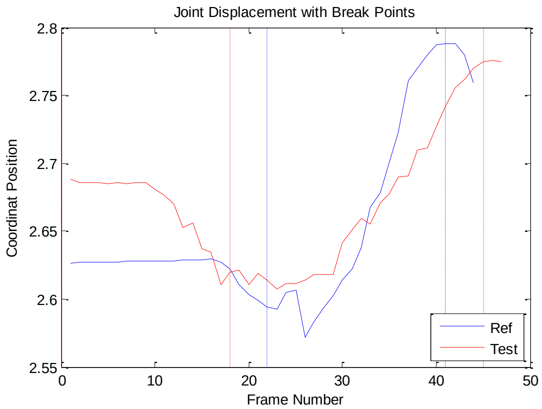

Fig. 5. Break Point Detection

It can be said that the action is similar but only the temporal shifts are observed.F_i^{td}The time shifts for the *i-th* breakpoint represent the errors in terms of amplitude for the *i-th* breakpoint. These values can be calculated for each time interval to find total slip and total error.

$$F_i^{ad} = F_{i,t}^{bp_x} - F_{i,t}^{bp_y} \tag{5}$$

$$F_i^{td} = F_{i,t}^{bpp_x} - F_{i,t}^{bpp_y} \tag{6}$$

x represents the person performing the reference movement, y the person performing the test movement. Direct comparison of acquired error information may not always yield accurate results. All breakpoints must match one another in order to give correct results. An erroneous hand movement, either at the beginning of the movement or at the time of its actualization, may cause the entire movement to be marked incorrectly. In such cases, we can provide the correct operation of the system with the DTW method. For this, it is tried to find the minimum distance between the series of reference breakpoints and the series of test breakpoints. By calculating the total cost computation according to Bellman's principle in total cost computation [21], the cost between two movements is found.

$$D_{i,j} = \min \begin{Bmatrix} F_{i,t+1}^{bp_x} - F_{i,t}^{bp_y} \\ F_{i,t+1}^{bp_x} - F_{i,t+1}^{bp_y} \\ F_{i,t}^{bp_x} - F_{i,t+1}^{bp_y} \end{Bmatrix} + \mathrm{d}(\mathrm{i}_k, \mathrm{j}_k) \tag{7}$$

Figure 6 shows two joint trajectory and local keyframes. For the reference action, at the keyframes in the joint trajectory t_1, at the fracture point t_3, instantaneously for the test action. The time shift t_{diff} can be calculated by taking the difference.

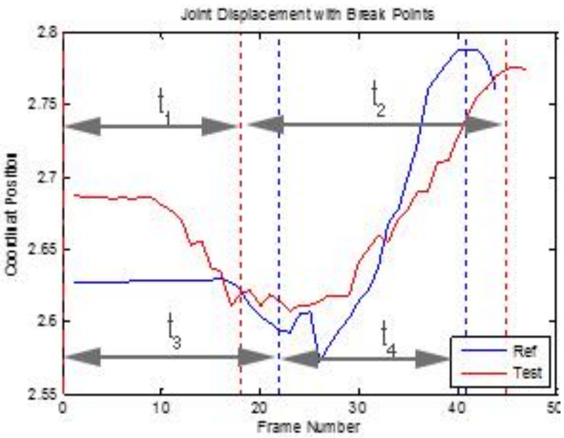

Fig. 6. Temporal shifts and amplitude difference

Not all joints move equally during an action. So while classifying it will not have an effect at the same time. To do this, each insertion must be assigned a weight value. In this study, as a function of weight, the ratio of the total change value of each joint to the total change value of all movements is found. Thus, each joint has its effect on the change value.

$$F_i^w = \frac{\sum_{t=1}^{n} \left| F_{i,t}^{d} \right|}{\sum_{i=1}^{i=n} \sum_{t=1}^{tn} \left| F_{i,t}^{d} \right|} \tag{8}$$

$$F^{err} = \sum_{i=1}^{n} D_{i,j} * F_i^w \tag{9}$$

The error value we will use for classification is *F* err. When classifying the test movement, the error values of all reference movements are calculated. Then classification is done using the k-nn algorithm.

E. Feedback

At the same time, the time shifts and joint faults for the detected class are presented to the user as feedback. Time shifts are temporal differences that occur as a result of fast or slow realization of action. Joint defects are the points that the joints have to reach or can not reach exactly.

III. Experimental Study

The developed method has been tested on a sample dataset. Test results are given in this section.The dataset prepared by Celebi et al. Was used as the data set for testing the work. This data set is a time series data set containing 224 motion information. In total there are 8 different gesture and 28 time series of each gesture. As in the original study, of these 28 time series, 8 were used for training and 20 were used for testing [12]. The gesture names are given in Table 1.

TABLE. I. MOVEMENTS IN THE DATASET

Number of Gesture	Name of Gesture
1	Left Hand Pull Down
2	Left Hand Pull Up
3	Left Hand Swipe Right
4	Left Hand Wave
5	Right Hand Pull Down
6	Right Hand Pull Up
7	Right Hand Swipe Right
8	Right Hand Wave

A. Results

In the developed application, action recognition is performed using joint motion keyframes. There are 8 different gesture in the dataset used as input in the study and 28 time series for each move [16]. The confusion matrix of the developed application over the dataset is given in Table 2.

TABLE. II. CONFUSION MATRIX

Number Of Gesture	1	2	3	4	5	6	7	8
1	100	0	0	0	0	0	0	0
2	0	100	0	0	0	0	0	0
3	0	0	100	0	0	0	0	0
4	0	0	0	100	0	0	0	0
5	0	0	0	0	100	0	0	0
6	0	0	0	0	0	100	0	0
7	0	0	0	0	0	0	100	0
8	0	0	0	0	0	0	0	100

In this study, a highly successful motion recognition application was performed using the detection of keyframes in the trajectory of all body joints. The success percentages of the studies performed with the literature studies that using same dataset are given in Table 3.

TABLE. III. SUCCESS PERCENTAGES

Method	Success
DTW	%84.41 [22]
State Of Art	%86.56
Weighted DTW	%97.13 [22]
WDTW With Keyframes	%100

IV. CONCLUSION

In this study, human action recognition system was developed using keyframes which constitute an action. The developed system has been tested on a sample dataset and found to work with high accuracy. keyframes are defined as the distinguishing feature for classifying actions. Also every joint is weighted and its effect is rearranged on the result. It has a high working speed thanks to the keyframes features it uses. The size of the features stored in the database has decreased in this respect. High operating speed has been achieved with few features. Therefore it is suitable for real time working. The system is robust against the noise generated by the user during the movement. Compared to other studies, high success was achieved. Because of highly discriminative feature vectors. All movements in the given data set are 100% successfully classified.

ACKNOWLEDGMENT

This study was supported by the Scientific Research Coordination Unit of Karabuk University under Grant KBÜ-BAP-17-DR-170.

REFERENCES

[1] Miranda, L., Vieira, T., Martinez, D., Lewiner, T., Vieira, A. W., & Campos, M. F. M. (2012). Real-time gesture recognition from depth data through key poses learning and decision forests. Brazilian Symposium of Computer Graphic and Image Processing, 268–275.

[2] L. Xia, C.-C. Chen, J. Aggarwal, View invariant human action recognition using histograms of 3D joints, in: Proceedings of Computer Vision and Pattern Recognition Workshops (CVPRW), IEEE, Providence, Rhode Island, 2012, pp. 20–27, http://dx.doi.org/10.1109/CVPRW.2012.6239233

[3] Chen, L., Wei, H., & Ferryman, J. (2013). A survey of human motion analysis using depth imagery. Pattern Recognition Letters, 34(15), 1995–2006.

[4] J. K. Aggarwal and M. S. Ryoo. Human activity analysis: A review. In ACM Computing Surveys, 2011.

[5] Mankoff, K. D. and Russo, T. A. (2013), The Kinect: a low-cost, high-resolution, short-range 3D camera. Earth Surf. Process. Landforms, 38: 926–936. doi:10.1002/esp.3332

[6] Raptis, M., Kirovski, D., & Hoppe, H. (2011). Real-Time Classification of Dance Gestures from Skeleton Animation. ACMSIGGRAPH Symposium on Computer Animation.

[7] Rosa-pujaz, A., Barbancho, I., Tard, L. J., &Barbancho, A. M. (2016). Fast-gesture recognition and classification using Kinect : an application for a virtual reality drumkit, 8137–8164. http://doi.org/10.1007/s11042-015-2729-8

[8] S.Z. Masood, C. Ellis, M.F. Tappen, J.J. LaViola, R. Sukthankar, Exploring the trade-off between accuracy and observational latency in action recognition, Int. J. Comput. Vis. 101 (3) (2013) 420–436, http://dx.doi.org/10.1007/ s11263-012-0550-7.

[9] X. Yang, Y. Tian, Eigenjoints-based action recognition using Naive-Bayes- Nearest-Neighbor, in: Proceedings of Computer Vision and Pattern Recog- nitionWorkshops (CVPRW), IEEE, Providence, Rhode Island, 2012, pp. 14–19, http://dx.doi.org/10.1109/CVPRW.2012.6239232.

[10] B. Li, M. Ayazoglu, T. Mao, O.I. Camps, M. Sznaier, Activity recognition using dynamic subspace angles, in: Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Colorado Springs, 2011, pp. 3193–3200, http://dx.doi.org/10.1109/CVPR.2011.5995672.

[11] N. Raman, S.J. Maybank, Action classification using a discriminative multi- level HDP-HMM, Neurocomputing 154 (2015): 149-161.

[12] Celebi, S., Aydin, A. S., Temiz, T. T., &Arici, T. (2013, February). Gesture Recognition using Skeleton Data with Weighted Dynamic Time Warping. In VISAPP (1) (pp. 620-625).

[13] C. Schuldt, I. Laptev, and B. Caputo, “Recognizing human actions: A local SVM approach,” in Proc. Int. Conf. Pattern Recognit., 2004, pp.32–36.

[14] B. Delachaux, J. Rebetez, A. Perez-Uribe, H.F.S. Mejia, Indoor activity recognition by combining one-vs.-all neural network classifiers exploiting wear- able and depth sensors, in: Advances in Computational Intelligence. Lecture Notes in Computer Science, Springer, Tenerife - Puerto de la Cruz, Spain, 7903 (2013), pp. 216–223.

[15] Town, C. and Sinclair, D., A self-referential perceptual inference framework for video interpretation. In International Conference on Vision Systems. (2003)

[16] Yang, K., &Shahabi, C. (2004). A PCA-based similarity measure for multivariate time series. Proceedings of the 2nd ACM International

Workshop on Multimedia Databases - MMDB ’04, 65. http://doi.org/10.1145/1032604.1032616.

[17] Chen, Y.-L., Wu, X., Li, T., Cheng, J., Ou, Y., &Xu, M. (2016). Dimensionality Reduction of Data Sequences for Human Activity Recognition. Neurocomputing, 1–9. http://doi.org/http://dx.doi.org/10.1016/j.neucom.2015.11.126

[18] Alwani, A., Salih, A., & Youssef, C. (2016). recognition using modified spherical harmonics, 1–10. http://doi.org/10.1016/j.patrec.2016.05.032

[19] Pazhoumand-Dar, H., Lam, C. P., &Masek, M. (2015). Joint movement similarities for robust 3D action recognition using skeletal data. Journal of Visual Communication and Image Representation, 30, 10–21. http://doi.org/10.1016/j.jvcir.2015.03.002

[20] Ibanez, R., Soria, Alvaro, Teyseyre, A., & Campo, M. (2014). Easy gesture recognition for Kinect. Advances in Engineering Software, 76, 171–180. http://doi.org/10.1016/j.advengsoft.2014.07.005

[21] Bellman, R. (1954). The theory of dynamic programming. Bull. Amer. Math. Soc, 60(6):503–515

[22] Arici, T., Celebi, S., Aydin, A. S., & Temiz, T. T. (2014). Robust gesture recognition using feature pre-processing and weighted dynamic time warping. Multimedia Tools and Applications, 72(3), 3045-3062.

[23] Bashar, Syed Khairul, and Mohammed Imamul Hassan Bhuiyan. "Classification of motor imagery movements using multivariate empirical mode decomposition and short time Fourier transform based hybrid method." Engineering Science and Technology, an International Journal 19.3 (2016): 1457-1464.

[24] Ren, Z., Meng, J., & Yuan, J. (2011, December). Depth camera based hand gesture recognition and its applications in human-computer-interaction. In Information, Communications and Signal Processing (ICICS) 2011 8th International Conference on (pp. 1-5). IEEE.