Human Gesture Recognition Using Kinect Camera

Orasa Patsadu, Chakarida Nukoolkit and Bunthit Watanapa
School of Information Technology
King Mongkut's University of Technology Thonburi
Bangkok, Thailand
54500701@st.sit.kmutt.ac.th
{chakarida | bunthit }@sit.kmutt.ac.th

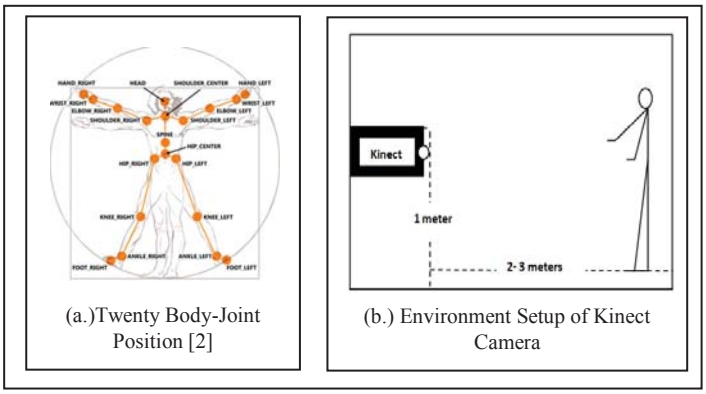
***Abstract*—In this paper, we propose a comparison of human gesture recognition using data mining classification methods in video streaming. In particular, we are interested in a specific stream of vector of twenty body-joint positions which are representative of the human body captured by Kinect camera. The recognized gesture patterns of the study are stand, sit down, and lie down. Classification methods chosen for comparison study are backpropagation neural network, support vector machine, decision tree, and naive Bayes. Experimental results have shown that the backpropagation neural network method outperforms other classification methods and can achieve recognition with 100% accuracy. Moreover, the average accuracy of all classification methods used in this study is 93.72%, which confirms the high potential of using the Kinect camera in human body recognition applications. Our future work will use the knowledge obtained from these classifiers in time series analysis of gesture sequence for detecting fall motion in a smart home system.**

Keywords- Kinect Camera; Human Gesture Recognition; Classification Methods; Body-Joint Positions; Video Streaming;

I. INTRODUCTION

Human gesture recognition [1] is defined as "a gesture as a human body movement. Human gesture is a non-vocal communication, used instead of or in combination with a verbal communication, intended to express meaning. It may be combined with the hands, arms or body, and also can be a movement of the head, face and eyes". Human gesture can be called several names [3, 4, 6, 7, 11], namely, human pattern, human posture, human pose, and human behavior. Human gesture recognition from video sequences has been heavily studied because of important applications to enhance monitoring of patients for fall motion detection, surveillance systems, motion analysis in sports, and human behavior analysis. Human gesture recognition may include standing, lying, bending, sitting, walking, and side walking, crouching, jumping, uphill and downhill actions. Our research focuses on a set of three gestures: stand, sit down, and lie down, to be a knowledge base of a smart home system which monitors and detects the fall motion of the elderly or hospital patients.

In this paper, we perform a variety of data mining classification methods and compare the performance based on classification accuracy in recognition of human gestures. The input data are streams of vectors of twenty body-joint positions obtained by standard Application Programming Interface (API) of the Kinect Software Development Kit (SDK). These joints represent the human body captured by Kinect camera [2] as shown in Figure 1 (a.). This study focuses on use of the Kinect camera because it is the latest consumer market gaming camera which is affordable (less than $150) and highly practical. The classification methods chosen for our research were based on previous research literature. They are backpropagation neural network (BPNN), support vector machine (SVM), decision tree (DT), and naive Bayes (NB).



(a.)Twenty Body-Joint Position [2]

(b.) Environment Setup of Kinect Camera

Figure 1. Twenty Body-Joint Positions [2] and Kinect Camera Setup

This paper is organized as follows: Section II Related Work; Section III Proposed System Overview; Section IV Training and Testing Datasets; Section V Preprocessing; Section VI Methodology; Section VII Experiment and Results; and Section VIII Conclusion and Future Work.

II. RELATED WORK

In the past, human gesture recognition has been based on computer vision and video-based techniques, in which the performance of recognition depends mainly on light conditions, shadow, and camera angles. However, the system performance using a single camera may suffer in the case of obstruction of subjects. Therefore, further research used multi-camera to solve occlusion. Wu and Aghajan [3] described a method of human posture estimation in a multi-camera network by using the concept of an opportunistic fusion framework, which is

978-1-4673-1921-8/12/$31.00 ©2012 IEEE

composed of three dimensions, space, time, and feature levels to obtain a 3D human skeleton.

Furthermore, Gavrila and Davis [4] described tracking and recognition of human movement on a 3D model from multi-view real images. The result of this research is 3D pose-recovery and movement classification.

Another approach of human gesture recognition uses an accelerometer, which gives the most accurate measurements, but this approach is not practical for everyday use, because it is intrusive and users may forget to wear it. Wu et al [5] presented an accelerometer-based gesture recognition approach using Frame-based Descriptor and multi-class SVM (FDSVM).

Existing classification methods used to recognize human gestures are BPNN, SVM, decision tree, and naive Bayes. Cohen and Li [6] classified body posture with SVM technique from a 3D visual-hull constructed from a set of silhouette input data. The system returns the classified human body postures in the form of thumbnail images.

Cheng Mo et al. [7] proposed a human behavior analysis system which recognizes human postures as walking, bending down, and sitting. A multicategory SVM is used to classify human postures using a human star skeleton, angles of six sticks in the star skeleton, and object motion vectors.

Zhao and Liu [8] used a centroid-radii model as a shape descriptor to represent human posture in each frame. Non-linear SVM decision tree is used to recognize human postures: standing, lying, bending, sitting, walking, side walking, crouching, jumping, uphill, and downhill action.

C˘ernekova [9] proposed a video-based system from the silhouette features of a person such as standing, position of the fingertips, and position of the shoulders. A multi-class SVM is used to recognize input features and return 2D position of subject as output.

Corradini et al [10] proposed a new method to analyze and recognize human posture (such as stop, go left, go right, hello left, and hello right) using hybrid neural networks.

Kiran et al [11] presented a comparison of posture recognition between supervised and unsupervised learning algorithms, which consists of K-means, artificial neural network, self-organizing maps, and particle swarm optimization. The recognized postures are climbing, fighting, jumping, lying down, and pointing. The results show that the supervised learning algorithm performs better than the unsupervised learning algorithm.

Lately, Raptis et al [12] presented a real-time gesture classification system to recognize dance gesture based on motion of six teen main skeleton joints obtained from Kinect camera in real time. The result shows that the system has accuracy of 96.9% using the approximately 4-second record of skeletal motion.

III. PROPOSED SYSTEM OVERVIEW

The overview of the proposed system is shown in Figure 2. We take the real video image as a collection of capture frames to be still images, obtaining a set of vectors of twenty body-joint positions to recognize human gesture using various data mining classification methods (such as BPNN, SVM, decision tree, and naive Bayes). Then we compare the performance of each method to find the optimal classifier.

Our dataset includes three human gesture output labels, namely, stand, sit down, and lie down, in Figure 3. We establish an indoor environment setting with single Kinect camera to monitor static scenes. The camera is located on the plane of the body, with two approximate distance settings: 2 and 3 meters. The user stands in front of the camera. The camera is approximately 1 meter from the floor, as shown in Figure 1 (b.).

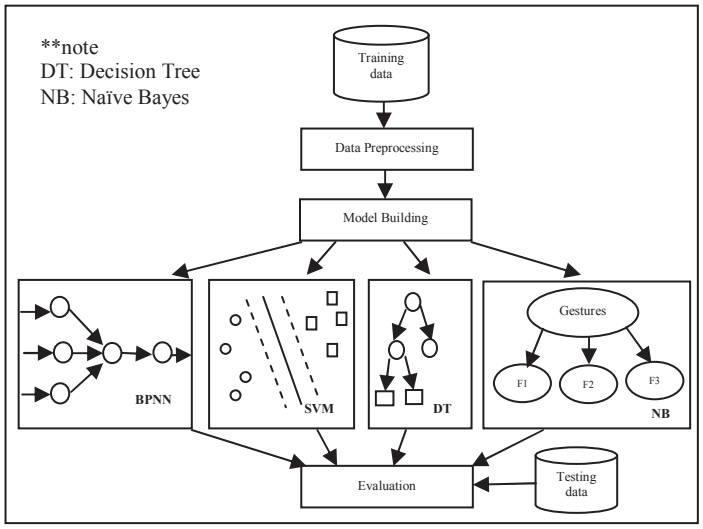


Figure 2. Overview of the proposed system

IV. TRAINING AND TESTING DATASETS

In the experiment, data mining classification methods, which consist of BPNN, SVM, decision tree and naive Bayes, were used to train the system for classification of three human gestures (stand, sit down, and lie down). Each frame of the video was represented by a row of the vector of twenty body-joint positions in (x, y, z).

There are six subjects, equal numbers of males and females, of various heights and weights. There are two camera distances (two and three meters). There are 7,200 and 3,600 records in the training and testing data sets respectively.

V. PREPROCESSING

This step is very important; we perform Z-score normalization to deal with parameters of different units and scales of body-joint positions.

In Z-score normalization [13], the values of an attribute "*A*" are normalized based on the mean and standard deviation of *A*. A value v of *A* is normalized to v' by computing:

$$v' = ((v - \overline{A}) / \sigma_A) \quad (1)$$

Where $\overline{A}$ and σ_A are the mean and the standard deviation of attribute *A*.

VI. METHODOLOGY

Several data mining classification methods are used to classify human gestures by using KNIME [14], which is a well-known open source data mining tool.

In this study, we selected four popular data mining classification methods: BPNN, SVM, decision tree, and naïve Bayes. These methods have been used in much of the previous research literature. We found that each approach is effective for human gesture recognition.

i.) Process of classification

Figure 2 demonstrates the proposed process of human gesture recognition using Kinect camera. Each classification method takes preprocessed input vectors of twenty body-joint positions as both training and testing data. The input data contains two distance settings of Kinect camera (two and three meters). There are 1,200 input vectors for each of the three human gesture classes in input data. Each distance setting contains 3,600 input vectors (x, y, z) of twenty body-joint positions as shown in Figure 3. This results in 7,200 input vectors in total for both camera distance settings. The testing data contains similar vector structure, and each class contains 1,200 vectors for both camera distance settings. Therefore, the output data contain 3,600 vectors in total. The classification result of each classifier is further illustrated as a set of confusion matrix as shown in Table 1.

ii.) Classification methods

a. Backpropagation Neural Network (BPNN)

BPNN [15, 16] is a multilayer feed forward neural network, which uses backpropagation algorithm in its learning. We use a multiclass neural network to predict class membership of human gestures (stand, sit down, and lie down). We are applying BPNN methods to inductively construct a model of data. There are three layers (input layer, hidden layer, and output layer) with 60, 10, and 3 nodes, respectively.

b. Support Vector Machine (SVM)

SVM [15, 17] is a promising new approach that can classify both linear and nonlinear data. Non-linear mapping is used to transform the training data within a higher dimension into a new dimension; linear mapping is used to search for a linear optimal line to separate hyperplanes. SVM consists of nodes used to train a support vector machine on the input data. It supports a number of different kernels (hyper tangent, polynomial, and radial basis function). The SVM learner supports multiple-class problems as well by computing the hyperplane between each class and the rest. In our study, we use SVM with polynomial kernel to classify and analyze regression of human gestures.

c. Decision Tree

Decision Tree [15] is used to classify data from class label, which yields output as a flow chart-like tree structure. In this research, a decision tree algorithm called CART is used based on its popularity in data mining research literature. In this study, the decision tree classifies human gestures as a set of internal nodes (decision nodes) and leaf nodes. Each leaf node shows a class outcome label. The constructed tree branches present outcomes of human gestures (stand, sit down, and lie down).

d. Naïve Bayes

Naïve Bayes [15] is a statistical classification which predicts class membership based on conditional probabilities. The nodes in a Bayesian model are created from given training data. Each node counts the number of rows per attribute value per class for nominal attributes and calculates the Gaussian distribution for numerical attributes.

VII. EXPERIMENT AND RESULTS

In this section, the comparison of classification methods used in recognizing human gesture is presented. Each classifier aims to recognize three different kinds of human gestures as shown in Figure 3.

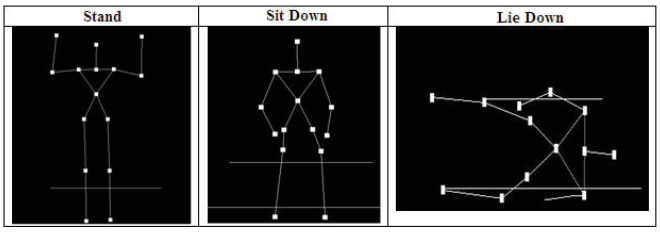


Figure 3. Human Gesture Recognition from Kinect Camera

In this experiment, we summarize the result of each classification method as shown below:

- Backpropagation Neural Network (BPNN)

From Table 1, the results show that this method is the most accurate classification method, with its accuracy of 100%. Therefore this method is a strong candidate for the classifier to be used to recognize human gestures with Kinect camera. BPNN was able to achieve 100% recognition accuracy for all the human gesture categories. The most likely reason for this method's success is that the datasets of images used for training and testing are free of excessive noise, and the gesture silhouette of the human is easily distinguished in the images.

- Support Vector Machine (SVM)

We have experimented with our SVM using several settings of kernel functions: a linear kernel, a quadratic polynomial kernel, and the radial basis function kernel. In this study, the polynomial kernel provides the most accurate result of classification.

From Table 1, the results show that the SVM classifier achieves 99.75% accuracy. We found that this method is also an effective and proper candidate for gesture recognition despite being slightly less accurate than the neural network

classifier. This method can separate the testing data into three categories with extreme accuracy. This finding confirms the fact that SVM models are closely related to neural networks.

- Decision Tree

We have used a CART decision tree with the *gini* index for tree quality measure to recognize human gestures. Since our data are normalized z-score numeric values, each node of our decision tree split the trees into two outcomes. Each node represents each human gesture threshold for each body-joint position. Figure 4 represents an example of sit down internal node using *x* axis of left hand as a condition of the decision.

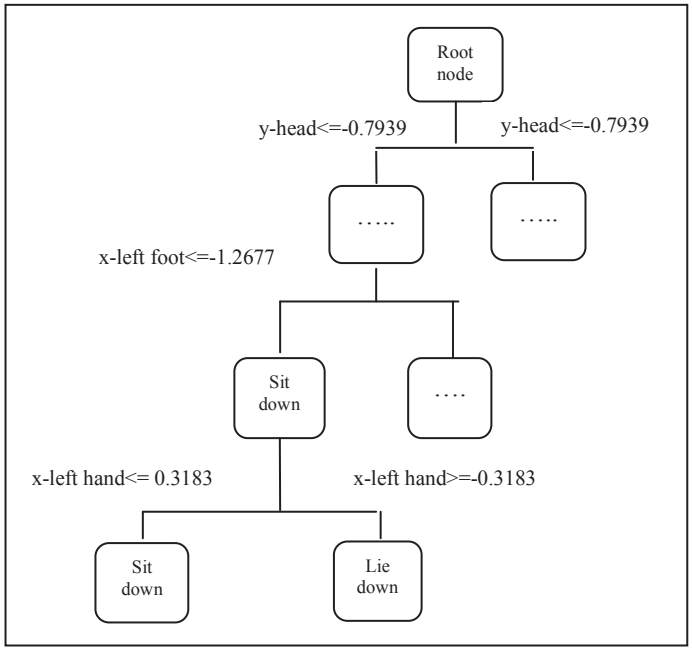


Figure 4. An Example of Sit Down Node from CART Decision Tree

From Table 1, the result of this decision tree is represented by confusion matrix, and the result shows that this method can classify with 93.19% accuracy. The small classification errors occur when there is ambiguity between the stand gesture of a short subject and the sit down gesture of a tall subject.

- Naïve Bayes

We have used naïve Bayes to recognize human gestures. We used KNIME default parameters for this method.

From Table 1, the naïve Bayes classifier performs with 81.94% accuracy. The error of classification comes from naïve Bayes's underlying assumption of feature-independent probabilities, which is not the case in this study. From the experiment, we found that this method might not be a good candidate method for human gesture recognition because of its lowest accuracy rate among the examined classification methods. The naïve Bayes classifier poorly handles the ambiguity when the tall and short subject sit down and stand up, respectively. The tall subject's silhouette height when sitting down is similar to the silhouette height of the short subject when standing.

From the experiment, there are four kinds of data mining classification methods (namely BPNN, SVM, decision tree, and naïve Bayes), which have been used to classify human gestures based on input data captured with Kinect camera. The performance of each classification method is compared as seen in Table 1. We found that each method has different accuracy dependent on the principles of each method. BPNN has the highest accuracy among all methods. This method shows the strongest potential to be used in recognition of human gestures.

In addition, from Table 1, both BPNN and SVM are highly accurate, with only 0.25% error in SVM, where in BPNN, there is zero error. Thus, we are certain that both methods are suitable and effective for use in future applications which rely on human gesture recognition, such as fall motion detection.

In the next step of our study, we plan to use both classification methods to enhance the more complicated recognition in time-series of video image of human gestures obtained from Kinect camera.

VIII. CONCLUSION AND FUTURE WORK

In this paper, we present human gesture recognition using Kinect camera. Our dataset consists of automatically extracted input vectors of twenty body-joint positions. The system successfully classifies input frames into three class gestures (stand, sit down, and lie down). In our experiment, the camera is located on the plane of the body in two distance settings (2-3 meters). Data mining classification methods, which include BPNN, SVM, decision tree, and naïve Bayes, are investigated for gesture recognition. In the experiment, the results indicate that BPNN shows superior performance compared to other classification methods, and can recognize human gestures with 100% accuracy. In addition, the average accuracy of all classification methods in both camera distances is 93.72%, which confirms the effectiveness of using the Kinect camera in human body recognition applications. In the near future, we plan to augment the knowledge of this recognition in time series analysis of a stream of human gestures in fall motion detection for practical and affordable smart home applications.

ACKNOWLEDGMENT

This work was supported by the Higher Education Research Promotion and National Research University Project of Thailand , Office of the Higher Education Commission; School of Information Technology, King Mongkut's University of Technology Thonburi; Rajamangala University of Technology Krungthep; and Mr.Anthony French for English proof-reading.

REFERENCES

[1] Mohamed B´echa KAˆA NICHE, "Human Gesture Recognition", Thesis (PhD), Universit´e de Nice - Sophia Antipolis – UFR Sciences, October, 2009.

[2] Microsoft Corporation, “Kinect for Xbox 360-Xbox.com”, [Online]. Available: http://www.xbox.com/en-GB/kinect/, [2011, May 12].

[3] Chen Wu and Hamid Aghajan, “Model-based Human Posture Estimation for Gesture Analysis in an Opportunistic Fusion Smart Camera Network”, Proceedings of International Conference on Advanced Video and Signal Based Surveillance, pp. 453 – 458, 2007.

[4] D. M. Gavrila and L. S. Davis, “Towards 3-D Model-Based Tracking and Recognition of Human Movement: a Multi-View Approach”, International Workshop on Automatic Face- and Gesture-Recognition”, pp.272-277, 1995.

[5] Jiahui Wu, Gang Pan, Daqing Zhang, Guande Qi, and Shijian Li, “Gesture Recognition with a 3-D Accelerometer”, Proceedings of the 6th International Conference on Ubiquitous Intelligence and Computing, pp. 25–38, 2009.

[6] Isaac Cohen, Hongxia Li, “Inference of Human Postures by Classification of 3D Human Body Shape”, International Workshop on Analysis and Modeling of Faces and Gestures, pp. 74 – 81, 2003.

[7] Hao-Cheng Mo, Jin-Jang Leou, and Cheng-Shian Lin, “Human Behavior Analysis Using Multiple 2D Features and Multicategory Support Vector Machine”, Proceedings of the International MVA2009 IAPR Conference on Machine Vision Applications, May 20-22, pp.46-49, 2009.

[8] Haiyong Zhao and Zhijing Liu, “Human Action Recognition Based on Non-linear SVM Decision Tree”, Journal of Computational Information Systems, pp.2461-2468, 2011.

[9] Z. C˘ernekova´, N. Nikolaidis and I. Pitas, “Single Camera Pointing Gesture Recognition using Spatial Features and Support Vector Machines”, Proceedings of the 15th European Signal Processing Conference (EUSIPCO-2007), pp.130-134, 2007.

[10] Andrea Corradini, Hans-Joachim Boehme, Horst-Michael Gross, “Visual-based Posture Recognition using Hybrid Neural Networks”, Proceedings of European Symposium on Artificial Neural Networks Bruges (Belgium) (ESANN'1999), 21-23 April, pp.81-86, 1999.

[11] Maleeha Kiran, Chee Seng Chan, Weng Kin Lai, Kyaw Kyaw Hitke Ali, and Othman Khalifa, “A Comparison of Posture Recognition using Supervised and Unsupervised Learning Algorithms”, Proceedings of International Conference on Computer and Communication Engineering (ICCCE), pp.1-6, 2010.

[12] Michalis Raptis, Darko Kirovski, Hugues Hoppe, “Real-Time Classification of Dance Gestures from Skeleton Animation”, Proceedings of the 2011 ACM SIGGRAPH/Eurographics Symposium on Computer Animation, 2011.

[13] Luai Al Shalabi, Zyad Shaaban and Basel Kasasbeh, “Data Mining: A Preprocessing Engine”, Journal of Computer Science 2 (9): pp.735-739, 2006.

[14] KNIME, “KNIME - Professional Open-Source Software”, [Online]. Available: http://www.knime.org, [2011, September 20].

[15] Jiawei Han and Micheline Kamber, “Data Mining Concepts and Techniques”, Morgan Kaufmann Publishers, Second Edition, 2006.

[16] Simon Haykin, “Neural Networks and Learning Machines”, Prentice Hall, Third Edition, 2008.

[17] Nello Cristianini and John Shawe-Taylor, “An Introduction to Support Vector Machines and Other Kernel-based Learning Methods”, Cambridge University Press, First Edition, 2000.

TABLE I CONFUSION MATRIX OF FOUR CLASSIFIERS

Prediction / Actual	Back Propagation Neural Network			Support Vector Machine			Decision Tree			Naïve Bayes		
	Stand	Sit down	Lie down	Stand	Sit down	Lie down	Stand	Sit down	Lie down	Stand	Sit down	Lie down
Stand	1200	0	0	1200	0	0	1133	67	0	656	362	182
Sit down	0	1200	0	0	1200	0	123	1077	0	0	1094	106
Lie down	0	0	1200	0	9	1191	0	55	1145	0	0	1200