Comparison of Head Movement Recognition Algorithms in Immersive Virtual Reality Using Educative Mobile Application

Nehemia Sugianto[1]
and Elizabeth Irenne Yuwono[2]
Ciputra University, Indonesia
[1]nsugianto@ciputra.ac.id
[2]eirene@ciputra.ac.id

Abstract - **Virtual reality has been implemented in many fields recently escpecially in education because its capability to produce a virtual world and take users to experience in different level with lower cost. The users will interact with the virtual world using their body or some parts of body such us head, hand, or voice. The problem of recognition accuracy level is still a challenging problem. This research is focused on comparing head movement recognition algorithms in a simple educative mobile application. Accelerometer sensor and RGB camera in Kinect are used to capture five basic head movements; nodding, shaking, looking up, looking down, tilting. Three different algorithms are used to recognize the movement; backpropagation neural network, dynamic time wrapping and convolutional neural network. The result reveals that accelerometer-based dynamic time wrapping method is the best method in recognizing the head movement with 80% accuracy level, followed by backpropagation neural network and convolutional neural network.**

Keywords - **Backpropagation Neural Network, Convolutional Neural Network, Dynamic Time Wrapping, Head Movement Recognition**

I. INTRODUCTION

Virtual reality could produce a virtual world which has many virtual objects and take users to interact using additional sensors such as Kinect, Leap Motion or Intel Real Sense to get experience in different level with lower cost. This technology could bring impossible experience (such as experiencing ancient world or dinosaurs world), high risk or high cost experience (such as experiencing deep water, outer space or operating human body) for various purposes; education or game [3]. Based on sense of presence, virtual reality has three types; non immersive virtual reality, semi immersive virtual reality and total immersive virtual reality. First type uses traditional display (such as personal computer or laptop) and simple controller (such as mouse, keyboard) or motion detection sensor. The cost is low but it has limited field of range and sense of immersive. Second type uses larger screen or multiple projector (which projected in 2D/3D view) and motion detection sensor. The cost is high and it has distortion problem but could give better immersive level. The third type uses specialized device called head-mounted display/HMD (such as Google Cardboard, Samsung Gear VR or Oculus Rift) to bring user to experience the virtual world fully. It has full range of view that makes user could explore the virtual world freely which gives best immersive level. The cost of device is various starting from cheap to expensive, depends on the materials and comfort level. In usage, the users will interact naturally with the virtual world in many ways using additional sensor. Users can interact using body movement, some parts of body (such as finger, hand or head movement) or voice. Sensor choices are determined by the interaction type.

A. Human Movement

Based on linguistic psychology research, human movements are classified into four types; conversation, controlling, manipulation and communication movement [8] which involving whole body or some parts of body. Head is the mostly used part of body in movement. There are some body movements that used in daily life [6]:

1. **Nodding:** This movement is used to address agreement or as persuasive tool in communication. This movement starts from moving head to up and down repeatedly.

2. **Shaking:** This movement is used to address rejection, disagreement or negative response in communication. This movement starts from moving head to left then right repeatedly.

3. **Looking up:** This movement is basic movement and used to give neutral response. This movement is moving head to up for a few moment.

4. **Looking down:** This movement is basic movement and used to give negative response, judge or being aggresive. This movement is moving head to down for a few moment.

5. **Tilting:** This movement is used to address tiredness or giving up. This movement is tilting head to left or right side.

B. Head Movement Recognition

Based on input signal, head movement recognition method is divided into four types; computer vision signal, acoustic signal, accelerometer or gyroscope signal and hybrid signal [8]. This research is focused on recognizing head movement using computer vision signal and accelerometer and gyroscope signal using head-mounted display. There are three algorithms used; classification using neural network with magnified gradient algorithm [4] dynamic time wrapping algorithm and convolutional neural network algorithm [2]. The input sign of first and second algorithm is accelerometer signal. Gyroscope signal is not used because not provided in most devices. The input of third algorithm is color image from RGB camera.

C. Accelerometer-Signal-Based Head Movement Features

These features are retrieved from accelerometer sensor provided in mobile device in certain time. The sensor calculates acceleration values againts earth gravity which representing device direction. The values are x acceleration, y acceleration and z acceleration.

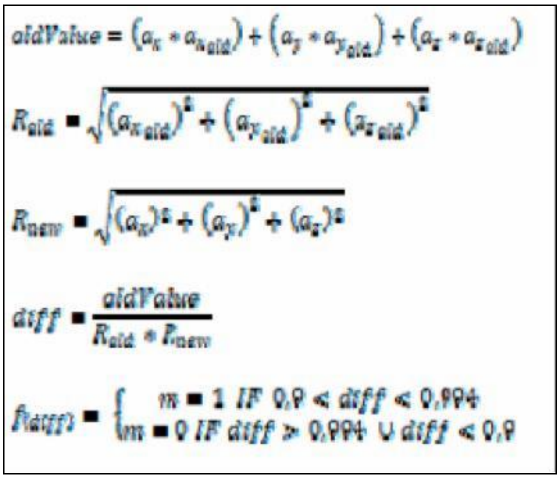
1. **X acceleration:** positive value equals to left direction, negative value equals to right direction.

2. **Y acceleration:** positive value equals to bottom direction, negative value equals to top direction.

3. **Z acceleration:** value will be A + 9.81 if the distance between sensor and ground level is becoming greater. A = Z acceleration value – earth gravity (-9.81 m/s^2)

4. **Z acceleration:** will be + 9.81 if the sensor is not moving (= 0 m/s^2 – (-)9.81 m/s^2)

Those values are converted into resultan value which represents device direction at current time. The formula to calculate resultan can be found in Fig. 1 [7].



$$oldValue = (a_x * a_{x_{old}}) + (a_y * a_{y_{old}}) + (a_z * a_{z_{old}})$$

$$R_{old} = \sqrt{(a_{x_{old}})^2 + (a_{y_{old}})^2 + (a_{z_{old}})^2}$$

$$R_{new} = \sqrt{(a_x)^2 + (a_y)^2 + (a_z)^2}$$

$$diff = \frac{oldValue}{R_{old} * R_{new}}$$

$$f(diff) = \begin{cases} m = 1 \; IF \; 0,9 < diff < 0,994 \\ m = 0 \; IF \; diff > 0,994 \cup diff < 0,9 \end{cases}$$

Fig. 1 Formula to Calculate Resultan Value

D. Computer Vision-Signal-Based Head Movement Features

The features are sequence of images of head in fixed sized, captured using RGB camera in certain time. For better head detection, head images are captured by localizing head area using head joints from Kinect sensor then normalized into fixed size. For better performance, the user must stand in front of sensor in certain range and the head must be segmented from the background image (background removal).

These features are retrieved from accelerometer sensor provided in mobile device in certain time. The sensor calculates acceleration values againts earth gravity which representing device direction. The values are x acceleration, y acceleration and z acceleration.

II. METHOD

This research consists of three stage; data retrieval, head movement recognition (pre-processing phase, feature extraction phase and head movement recognition) and data training.

A. Data Retrieval

Data is collected by four respondents (two males and two females) between 17-32 years old. Each respondent performs 15 repetitions for each movement using Samsung Gear VR. Number of data collected is 300 data sets (15 repetitions x 5 movements x 4 respondents). Data is recorded by using Samsung S7 Edge and RGB camera of Kinect placed in front of the user. 200 correct data sets is selected randomly from 300 data sets to eliminate incorrect movements (such as wrong answers or incomplete head movements). 80% from data sets collected is used for training phase, 20% from data sets is used for testing phase. To make respondents perform movement naturally, each respondent is asked to answer some questions using 5 movement choices until produces expected natural movements in certain number. If the questioner runs out of the questions, then they can improvise the questions. This strategy is used to avoid excessive movements. Unfortunately, this strategy is only for nooding and shaking movement. For the other movements, the questioners will give instructions to the respondents to perform the movement. The list of questions can be found at Table I.

TABLE I
LIST OF QUESTIONS FOR FIRST AND SECOND MOVEMENTS

Movement	Question
Nodding/ shaking	• Are you male? • Are you female? • Are you married? • Do you already have children? • Are you working? • Do you come from Surabaya? • Do you like fried chicken? • Do you like fried rice?

B. Head Movement Recognition Using Backpropagation Neural Network

Duration of each movement is about 1-2 seconds. Each movement is divided into 10 segments equally. Each segment has 3 acceleration values (X acceleration, Y acceleration and Z acceleration). For each segment, resultan is calculated. The input data of this method is difference values between current resultan and previous resultan (r_i-r_{i-1}, r_{i-1}-r_{i-2}, r_{i-2}-r_{i-3} …).

This method uses backpropagation neural network. This network has 10 neurons in input layer, 10 neurons in hidden layer and 5 neurons in output layer. Number of neurons in hidden layer is calculated from ⅔ x number of neurons in input and output layers. Neurons in output layer represent as number of head movement recognition classes. This network uses sigmoid biner activation function for each layer and threshold value = 0.7. This network uses Nguyen algorithm to initiate network weights. This network is trained using backpropagation method with MSE = 0.001, learning rate = 0.2, momentum = 0.5 and maximum epoch = 1,000 epochs. Those values are determined by experiments.

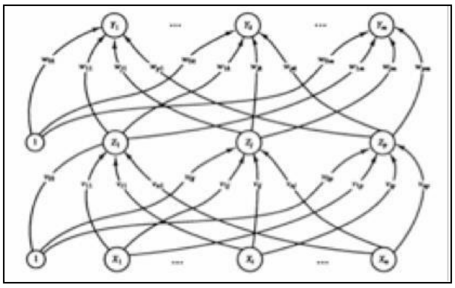


Fig. 2 Backpropagation Neural Network

C. Head Movement Recognition Using Dynamic Time Wrapping

The input data of this method is same with input data of previous method; difference values between current resultan and previous resultan (r_i-r_{i-1}, r_{i-1}-r_{i-2}, r_{i-2}-r_{i-3} …). Duration of each movement is about 1-2 seconds. Each movement is divided into 10 segments equally. Each segment has 3 acceleration values (X acceleration, Y acceleration and Z acceleration). For each segment, resultan is calculated. The input data of this method is difference values between current resultan and previous resultan (r_i-r_{i-1}, r_{i-1}-r_{i-2}, r_{i-2}-r_{i-3}…).

D. Head Movement Recognition Using Convolutional Neural Network

Colour does not affect accuracy level in recognizing head movement. Colour only affects in recognizing human face.

Therefore, the input images must be converted into grayscale images for faster computation and lower computation cost.

Duration of each movement is about 1-2 seconds. Each movement is divided into 10 frames equally. The input data of this method is sequence of 10 grayscale images of head in fixed sized (150 pixels x 150 pixels).

This network has two main parts; feature extraction part and classification part. First part has function to extract features from grayscale head images. This part has 4 convolution layers (consists of max pooling operator, two fully-connected layers, output softmax layer). Second part has function to classify features into output classes using backpropagation neural network. This network has 10 neurons in input layer, 10 neurons in hidden layer and 5 neurons in output layer. Number of neurons in hidden layer is calculated from ⅔ x number of neurons in input and output layers. This network uses sigmoid biner activation function for each layer and threshold value = 0.7. This network uses Nguyen algorithm to initiate network weights. This network is trained using backpropagation method with MSE = 0.001, learning rate = 0.2, momentum = 0.5 and maximum epoch = 1,000 epochs. Those values are determined by experiments.

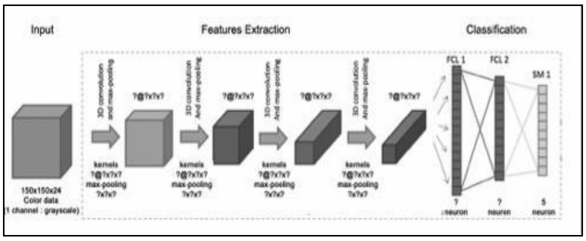

Input
Features Extraction
Classification
3D convolution and max-pooling
3D convolution And max-pooling
3D convolution And max-pooling
3D convolution And max-pooling
?@?x?x?
?@?x?x?
?@?x?x?
?@?x?x?
kernels ?@?x?x? max-pooling ?x?x?
kernels ?@?x?x? max-pooling ?x?x?
kernels ?@?x?x? max-pooling ?x?x?
kernels ?@?x?x? max-pooling ?x?x?
150x150x24 Color data (1 channel : grayscale)
FCL 1
FCL 2
SM 1
? neuron
? neuron
5 neuron

Fig. 3 Convolutional Neural Network

III. RESULT AND DISCUSSION

After the recognition systems are trained, the systems are tested in two steps to get recognition accuracy level for each algorithm; using training data (80%) and testing data (20%).

A. Head Movement Recognition Result Using Backpropagation Neural Network

Using training data sets, this method could recognize fifth movement (tilting) with 90.63% accuracy level, 87.50% accuracy level for third movement (looking up), 84.38% for fourth movement (looking down), 75.00% for first movement (nodding) and then followed by 71.88% for second movement (shaking). The average movement recognition of this method is 81.88%.

TABLE II
HEAD MOVEMENT RECOGNITION RESULT USING BACKPROPAGATION NEURAL NETWORK (USING TRAINING DATA SETS)

Move-ment	Movement					
	1	2	3	4	5	Failed
1	24	3	3	2	0	8
2	4	23	0	2	3	9
3	0	4	28	0	0	4
4	2	2	1	27	0	5
5	2	0	0	1	29	3
Accuracy level of first movement						: 75.00%
Accuracy level of second movement						: 41.88%
Accuracy level of third movement						: 87.50%
Accuracy level of fourth movement						: 84.38%
Accuracy level of fifth movement						: 90.63%
Average accuracy level of all movement						: 81.88%

Using testing data sets, this method could recognize fourth movement (looking down) with 90.63% accuracy level, 75.50% accuracy level for third (looking up) and fifth movement (tilting), and then followed by 62.50% for first (nodding) and second movement (shaking). The average movement recognition of this method is 72.50%.

TABLE III
HEAD MOVEMENT RECOGNITION RESULT USING BACKPROPAGATION NEURAL NETWORK (USING TESTING DATA SETS)

Move-ment	Movement					
	1	2	3	4	5	Failed
1	5	1	2	0	0	3
2	1	5	0	1	1	3
3	0	2	6	0	0	2
4	0	0	0	7	1	1
5	2	0	0	1	6	2
Accuracy level of first movement	: 62.50%					
Accuracy level of second movement	: 62.50%					
Accuracy level of third movement	: 75.00%					
Accuracy level of fourth movement	: 87.50%					
Accuracy level of fifth movement	: 75.00%					
Average accuracy level of all movement	: 72.50%					

B. Head Movement Recognition Result Using Dynamic Time Wrapping

Using training data sets, this method could recognize fifth movement (tilting) with 90.63% accuracy level, 87.50% accuracy level for third (looking up) and fourth (looking down), 81.25% for first movement (nodding), and then followed by 78.13% for second movement (shaking). The average movement recognition of this method is 85.00%.

TABLE IV
HEAD MOVEMENT RECOGNITION RESULT USING DYNAMIC TIME WRAPPING (USING TRAINING DATA SETS)

Move-ment	Movement					
	1	2	3	4	5	Failed
1	26	1	3	2	0	6
2	3	25	0	1	3	7
3	0	4	28	0	0	4
4	1	2	1	28	0	4
5	2	0	0	1	29	3
Accuracy level of first movement	: 81.25%					
Accuracy level of second movement	: 78.13%					
Accuracy level of third movement	: 87.50%					
Accuracy level of fourth movement	: 87.50%					
Accuracy level of fifth movement	: 90.63%					
Average accuracy level of all movement	: 85.00%					

Using testing data sets, this method could recognize fourth movement (looking down) with 87.50% accuracy level, 75.00% accuracy level for second (shaking), third (looking up) and fifth movement (tilting), and then followed by 62.50% for first (nodding). The average movement recognition of this method is 75.00%.

TABLE V
HEAD MOVEMENT RECOGNITION RESULT USING DYNAMIC TIME WRAPPING (USING TESTING DATA SETS)

Move-ment	Movement					
	1	2	3	4	5	Failed
1	5	1	2	0	0	3
2	0	6	0	1	1	2
3	1	1	6	0	0	2
4	0	0	0	7	1	1
5	2	0	0	0	6	2
Accuracy level of first movement	: 62.50%					
Accuracy level of second movement	: 75.00%					
Accuracy level of third movement	: 75.00%					
Accuracy level of fourth movement	: 87.50%					
Accuracy level of fifth movement	: 75.00%					
Average accuracy level of all movement	: 75.00%					

C. Head Movement Recognition Result Using Convolutional Neural Network

Using training data sets, this method could recognize fifth movement (tilting) with 90.63% accuracy level, 84.38% accuracy level for fourth movement (looking down), 78.13% for third movement (looking up), 68.75% for first movement (nodding) and then followed by 65.63% for second movement (shaking). The average movement recognition of this method is 75.00%.

TABLE VI
HEAD MOVEMENT RECOGNITION RESULT USING CONVOLUTIONAL NEURAL NETWORK (USING TRAINING DATA SETS)

Move-ment	Movement					
	1	2	3	4	5	Failed
1	22	2	3	2	3	10
2	4	21	2	2	3	11
3	1	4	25	0	2	7
4	1	3	1	27	0	5
5	4	2	1	1	24	8
Accuracy level of first movement	: 68.75%					
Accuracy level of second movement	: 65.63%					
Accuracy level of third movement	: 78.13%					
Accuracy level of fourth movement	: 84.38%					
Accuracy level of fifth movement	: 90.63%					
Average accuracy level of all movement	: 75.00%					

Using testing data sets, this method could recognize fourth movement (looking down) with 75.00% accuracy level, 62.50% accuracy level for first (nodding), third (looking up) and fifth movement (tilting), and then followed by 50.00% for second movement (shaking). The average movement recognition of this method is 62.50%.

TABLE VII
HEAD MOVEMENT RECOGNITION RESULT USING CONVOLUTIONAL NEURAL NETWORK (USING TESTING DATA SETS)

Move-ment	Movement					
	1	**2**	**3**	**4**	**5**	**Failed**
1	5	1	2	0	0	3
2	1	4	0	1	2	4
3	0	3	5	0	0	3
4	0	0	1	6	1	2
5	2	0	0	1	5	3

Accuracy level of first movement	: 62.50%
Accuracy level of second movement	: 50.00%
Accuracy level of third movement	: 62.50%
Accuracy level of fourth movement	: 75.00%
Accuracy level of fifth movement	: 62.50%
Average accuracy level of all movement	: 62.50%

Among three algorithms, dynamic time wrapping has best performance in recognizing head movement with 80.00% average accuracy level, then followed backpropagation neural network with 77.19% average accuracy level and convolutional neural network with 68.65% average accuracy level. Based on experiment and observation, dynamic time wrapping and backpropagation neural network perform better caused by some reasons; 1) resultan feature could give more precise and complete information in pattern recognition because resultan represents direction, rather than grayscale head image, 2) resultan value is not interfered by user background but head image is interfered by user background easily, and 3) distance between user and Kinect sensor could affect the quality of head image (size and detail) which this could require much more training data to recognize head movement.

TABLE VIII
COMPARISON OF ACCURACY LEVEL BETWEEN THREE ALGORITHMS

Algorithm	Using training data	Using testing data	Avg.
Backpropagation Neural Network	81.88%	72.50%	77.19%
Dynamic Time Wrapping	85.00%	75.00%	80.00%
Convolutional Neural Network	74.80%	62.50%	68.65%

IV. CONCLUSIONS

The result shows that those algorithms could recognize head movement well enough for time series data; 80.00% accuracy level using dynamic time wrapping, followed by 77.19% accuracy level using backpropagation neural network and 68.65% accuracy level using convolutional neural network. Based on experiment and observation, feature selection has big influence in recognizing head movement. Resultan feature could give better performance rather than head image because interfered by complex background and needs larger training data stes.

REFERENCES

(Arranged in the order of citation in the same fashion as the case of Footnotes.)

[1] Bautista, M.A., Hernandez-Vela, A., Ponce, V., Perez-Sala, X., Baro, X., Pujol, O., Angulo, C., and Escalera, S. (2012). “Probability-based Dynamic Time Warping for Gesture Recognition on RGB-D Data”. International Conference on Pattern Recognition: International Workshop on Depth Image Analysis (WDIA) Vol. 7854. Tsukuba, Japan.

[2] Cheron, G., Laptev, I., and Schmid, C. (2015). “Pose-based Convolutional Neural Network Features for Action Recognition”.

[3] Eggarxou, D. (2007). “Teaching History Using a Virtual Reality Modelling Language Model of Erechtheum”. International Journal of Education and

Development Using Information and Communication Technology (IJEDICT) Vol. 3 (3) pp. 115-121.

[4] King, L.M., Nguyen, H.T., and Taylor, P.B. (2005). “Hands-free Head Movement Gesture Recognition Using Artificial Neural Networks and The Magnified Gradient Function”. Proceeding of 27th Annual Conference of Enggineering, Medical, and Biology pp. 2063-2066.

[5] Rahayfeh, A. and Faezipour, M. (2013). “Eye Tracking and Head Movement Detection: A State-of-Art Survey”. IEEE Journal of Translational Engineering in Health and Medicine. DOI 10.1109/JTEHM.2013.2289879.

[6] Toastmasters International. (2016). “Dimensions of Body Language”. Diakses dari <http://westsidetoastmasters.com/resources/book_of_body_language/chap11.html>. Accessed 6 October 2016.

[7] Tolle, H., Pinandito, A., Muhammad, A.E., and Arai, K. (2015). “Virtual Reality Game Controlled With User’s Head and Body Movement Detection Using Smartphone Sensors”. ARPN Journal of Engineering and Applied Sciences pp. 9776-9782.

[8] Wu, Y. and Huang, T. (1999). “Human Hand Modeling, Analysis and Animation in the Context of HCI, IEEE Intl Conf”. Image Processing.

[9] Zhao, Z., Wang, Y., and Fu, S. (2012). “Head Movement Recognition Based on Lucas-Kanade Algorithm”. Proceedingof International Conference CSSS pp. 2303-2306.