Jurnal Teknologi

Full paper

The Optimization of Interface Interactivity using Gesture Prediction Engine

Mahdi Babaei*, Wong Chee Onn, Lim Yan Peng

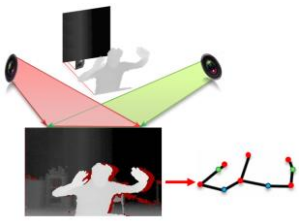
Faculty of Creative Multimedia, Multimedia University, Jalan Multimedia 63100 Cyberjaya Selangor Malaysia

*Corresponding author: Mahdi.babaei@gmail.com

Article history

Received :1 December 2013
Received in revised form :
10 January 2014
Accepted : 31 January 2014

Graphical abstract



Abstract

The primary objective of this project is to develop a gesture recognition engine for interactive interfaces using Microsoft Kinect device. A photo album is a sample of daily-use applications that is capable of having interactive interface. In this project there are features implemented to help users to view and edit their photos on the easier way. Although the 3D interface of photo album increases the reality and easy to use, simplicity of natural gestures which are recognizing by the gesture recognition engine eases the interaction. The contribution of this project is simultaneous work of a prediction and recognition engine. The algorithm benefits a Hidden Markov Model (HMM) state machine to record, update and calculate the occurrence probability of each gesture as a state in relation with previous states. It also aims to solve a major problem of interaction with the same applications which were their dependence on using devices physically and touch them directly. The optimized model had tested in an interactive digital space.

Keywords: Gesture recognition; photo album; gesture prediction; microsoft kinect; human computer interaction; smart interactivity

Abstrak

Objektif utama projek ini adalah untuk membangunkan satu sistem pengecaman gerak tubuh untuk antara-muka berinteraktif dengan menggunakan peralatan Microsoft Kinect. Satu sampel aplikasi yang dinamakan sebagai album foto akan mempamerkan aksi menggunakan antara-muka interaktif. Projek ini merangkumi ciri-ciri untuk membantu pengguna supaya pengguna dapat memanipulasi foto dengan berkesan. Walaupun antara-muka 3D album foto akan menyenangkan pengguna, aksi gerak tubuh semulajadi yang genang dikenal pasti akan menyenangkan interaksi. Sumbangan utama projek ini adalah enjin ramalan dan pengecaman gerak tubuh. Algoritma Hidden Markov Model (HMM) akan digunakan untuk merekod, mengemaskini dan pengiraan kebarangkalian setiap gerak tubuh yang akan dikaitkan dengan aksi gerak tubuh sebelumnya. Sistem ini juga bertujuan untuk menangani masalah interaksi dengan aplikasi yang sama di mana aplikasinya saling bergantung dengan peralatan fizikal dan sentuhan semulajadi. Model ini telah pun diuji dalam satu kawasan digital berinteraktif.

Kata kunci: Pengiktirafan isyarat; album foto; ramalan gerak isyarat; microsoft kinect; interaksi komputer manusia; interaktif pintar

1.0 INTRODUCTION

The interaction between human and computer machines was a big challenge since the earliest versions. The entrance of such electronic device as a part of daily-life increased the importance of optimized way and devices to establish such connection and perform the action. The human uses different kind of interactions like speech, gesture or gaze to communicate with each other (Chu and Cohen, 2005). Thus a human-computer interaction system that includes all those elements (to follow or simulate human gestures) can allow users to interact with computers in a more natural ways (Chu and Cohen, 2005). The history of HCI began with the popularity of personal computers and huge growth of technology and number of computer users. The quality of such interaction had been proposed when non-professional computer users constitute a big share of computer market, which they expect machines that anyone can use with even lack knowledge. A powerful machine that can work with natural easy gestures was the goal. This massive growth made the idea that good interaction with a computer can be considered as the most reliable, fast, easy and with lowest possible lag one. The popular hardware devices invented for such reason were named as keyboards or mice which can be categorized under mechanical devices.

Recognition of human body gestures aims to translate the human actions, filter meaningful ones, and extract useful data out of it. This data can be used to control machines or at least study on user behavior. Human hands (as a paired part of a person) have the most meaningful movements. These natural movements are ignored when we are dealing with a physical device like mice or keyboards.

One of the major factors that may affect the quality of an interaction with a computer system (from the user view) is the interface of the application that let the user interact with the system. The organization of accessing data, visual effects, friendly environment and easy interaction method consider as factors that an interactive application system must have. This project aims to recognize gestures to perform zoom in/out and rotation in an interactive photo album. In addition, the relation of each recognized gesture with previous ones needs to be discovered. It should also find the strength of such movements as it is needed to be used as the raw data for prediction engine. This engine calculates the probabilities of occurrence a certain action at next and estimate the strength and type.

This project contributes a natural interaction in digital space. There has been studies on gesture techniques and devices and finds the suitable one, in order to have a reliable and accurate prediction.

2.0 LITERATURE REVIEW

From late 1990s there are many researches done in the field of gesture recognition. Most of these attempts were based on a network of sensors attached to a human body (Takahashi and Kishino, 1991; Zimmerman *et al.*, 1987; Nanayakkara *et al.*, 2013). In some of them, sensors were placed exactly over human body skin of joints in order to track movement of joints. In some other only one sensor attached to each part of the body. Generally, all these approaches vary from mathematical models (i.e. Markov chain) to tools based on soft computing (Mitra and Acharya, 2007).

Gestures may contain symbolic information or affective one. These information can be received by tools ranging from statistical modeling, computer vision, image processing (Mitra and Acharya, 2007). They can get categorized into 3D models, appearance and skeletal (Pavlovic *et al.*, 1997) based ones.

Lately, most of the gesture recognition methodologies changed their path from sensor based to image processing and computer vision (Steps are shown in Figure 1). The best approaches in this field benefits (Mistry and Maes, 2009), a normal webcam device as a video capturing resource and an image processing engine in order to track color blobs and recognize body gesture. In June 2012, the industry was affected by a software development kit which was released by Microsoft and helped the programmers to drive Kinect. The Kinect is not only a device for a game console or a simple hardware available in the market. It goes further and gives application developers a great opportunity to be innovative because of a 3D sensing technology which is acquired to find a human body in a frame and recognize its movements (Oikonomidis *et al.*, 2011).

This project is based on skeletal type of recognition and observations come from an off-the-shelf Kinect sensor.

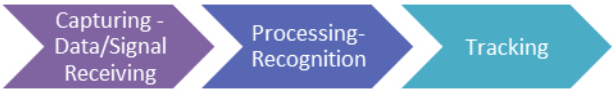


Figure 1 The general recognition process

2.1 The Gesture Recognition Engine

This system benefits from a combination of skeletal detection and image processing (Gonzalez and Woods, 1992). The software development kit (SDK) uses image processing in order to detect and position user movements and track user's skeleton in the skeleton engine (Webb and Ashley, 2012). It is obvious that the system can use the entire human body for interaction purpose but it is limited to hands because of these unique characteristics: 1) a posture: static finger configuration without hand. 2) a gesture: dynamic hand movements, with or without finger motion (Mitra and Acharya, 2007). These two factors made hands as the most expressive and frequently used human limb. Hand gesture recognition can show symbolic and emotional signs as well. It can be categorized into two major groups:

1- Sensor Base: there are number of sensors attached to the user's body in order to detect movements, rotation and angles (Mitra and Acharya, 2007).

2- Image Base: which is not using sensors and benefits from a video camera source and an image processing engine to detect body and track movements (Mitra and Acharya, 2007). Each of these methods can vary based on a number of cameras, speed, latency (Mitra and Acharya, 2007; Gavrila, 1999), 2D or 3D representation (Aggarwal and Cai, 1997). The most recent high standard device available in the market is Microsoft Kinect which we used for the project purpose. Since 2010 that Microsoft introduced this device (Sumar and Bainbridge-Smith, 2011) it had received lots of attention due to low price, high accuracy and great functionality.

Kinect is using an RGB camera (Khoshelham, 2011) and a depth infrared sensor combination (Oikonomidis *et al.*, 2011) to provide 3D information about objects in a scene. RGB and Depth sensor have 57 degree horizontal and 43 degree vertical field of view (Sumar and Bainbridge-Smith, 2011) with 30 frame rate per second. It has a powerful monochrome CMOS chip to capture 3 dimensions in low light condition (Hai *et al.*, 2011). The primary function of a Kinect device is to produce a depth map out of a scene (Webb and Ashley, 2012). Considering all the reasons discussed, Microsoft Kinect can be described as a device with high reliability rate even in research projects.

The algorithm of the developed application for this system had shown in Figure 2.

Figure 2 Application algorithm

Gesture recognition using Kinect begins by capturing a frame from the RGB-camera same as all other video capturing sources. Using depth frame from infrared camera the distances can be measured. The measurement of depth and mapping captured frame by infrared camera and RGB camera gives the value of depth for each individual point. Although the generated infrared points of a Kinect are limited, the extraction of human body from RGB frame limits the desired data to body joints only. A sample of extracted human body from depth frame based on mapping technique over RGB frame by Kinect is shown in Figure 3.

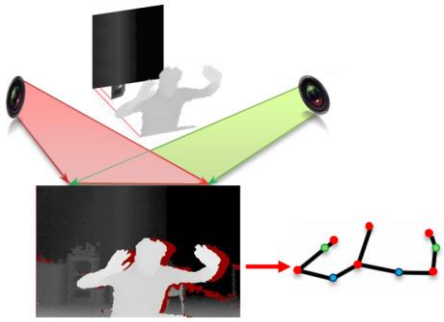


Figure 3 Kinect RGB and IR cameras, depth detection and skeleton extraction

In Figure 3 the skeleton joint point is extracted using sitting mode in Kinect latest SDK (version 1.7). In order to develop the gesture recognition engines, a complete joint extraction and tracking process is needed. As shown in Figure 4, the result of a sample movement tracking for a human by changing hands position to up and down is measured and recorded.

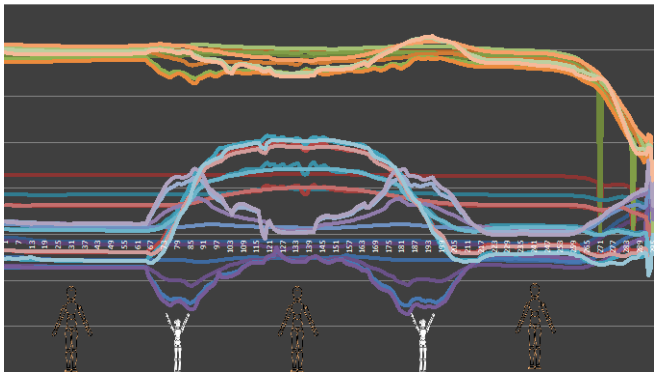


Figure 4 Moving hands received data analysis

As the diagram shows the movements of human body joints changes the positions numbers. The same diagram may result by tracking such movement in depth by changing hand position from the initial point to a second position more nearer to the camera. Tracking movement in each axis and comparison of moving path to a pre-defined one considers as the main approach toward recognition of gestures.

Mapping skeleton points from the depth frame on RGB frame is the next step in this algorithm. The final mapped hand skeleton joint points on screen will be the reference point of the mouse cursor. In Figure 7, red dots show the mapped point that is based on screen size. Although they will not be visible in final application, these are only to show the position.

The movement of right hand into a front position that is more nearer to the camera and less than a hand length considers as a click gesture. It is used in the main form that is an interactive interface for user to choose the photo. The swipe gesture is used to navigate between photos that are flipping in a 3D space. The swipe gesture is the movement of hand wrist and lower arm when elbow is fixed. In Figure 5, the initial wrist and lower arm is shown using red circles and the last point shown in green. The fixed elbow is in blue color.

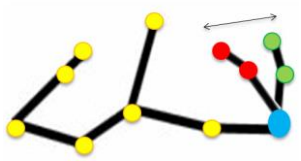


Figure 5 Swipe gesture detection using skeleton point

The recognition of rotation gesture is by tracking hands wrist while they are moving vertical on different direction. As Figure 6-A Shows the Rotation CCW is when user's right hand goes up and left hand down simultaneously. From another perspective a vertical scroll gesture by each of hands should be recognized on different direction and at the same time. In Figure 6-B, the scale and zoom gestures are recognized when the user's hands are moving in different direction horizontally and vertically.

A

B

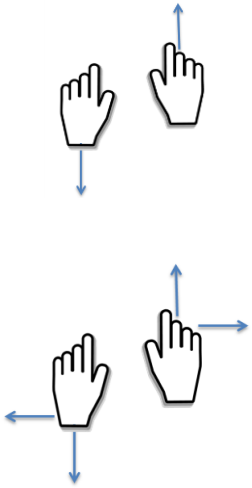


Figure 6 'A' zoom gesture, 'B' rotate gesture

A brief explanation of the gestures usage is:

a) Left/right navigation - This gesture enables user to browse the media collections.
b) Scaling - This gesture allows user to enlarge or shrink the media that is currently under selection.
c) Rotation - This gesture allows user to rotate the media that is currently selected by a maximum of 90 degrees for both clockwise and anti-clockwise.
d) Simultaneous Scaling and Rotation - The Scaling and Rotation gestures can be triggered at the same time.

Figure 7 Next or previous states, simultaneous rotate and scale state, rotation of selected photo

2.2 Prediction of Actions

The last objective of this project considers as the major factor to optimize the interactivity of an interface. It introduces a way beyond the use of interactive device and human gesture recognition to make the dialogue. It can predict the gesture that user is going to do as his next action by find and formulating each gesture as a state and find the relation between if there is any. The probability of current recognized gestures and occurrence of the next movement is highly depends on what user have done so far and iteration of such action. The probability of occurrence for each state calculates and changes in case of happening the rest of states. Application of Hidden Markov Model formula on initial points and the total of each row of states are shown in Figure 8.

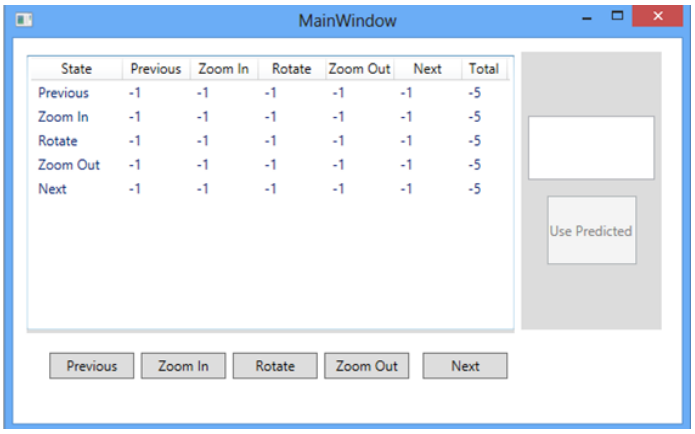


Figure 8 HMM state model initial point

This algorithm is a finite state machine and 5 states transition matrix. The machine states and the initial point of prediction engine in this system is shown in Figure 8. The gesture recognition engine designed to look for occurrence of a gesture. In order to perform this action there a chain of activity in order to look for each gesture (each state). The calculation of all necessary information which is shared between states occurred in the first state. Although this method works real-time, even in 1/30 of the second, the most time consuming part is recognition the first step due to having heavy calculation. The order of state checking is as shown in Figure 9.

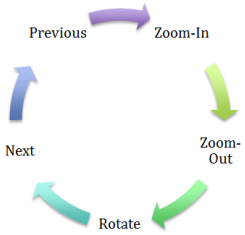


Figure 9 States order

The HMM (hidden Markov Model) (Yamato *et al.*, 1992; Rabiner, 1989) process governed by two processes which are finite number of states and random function for each of the states. Transition between states has transitioned probability and output probability (Rabiner, 1989). Figure 10 shows the states which existed in Figure 9.

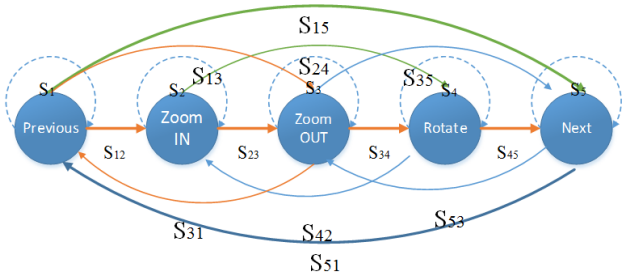


Figure 10 Possible transitions in this system

The key issues with this model are evaluated as follows:

1) The observed sequence was generated by model (Mitra and Acharya, 2007).

2) Estimation which is modeled adjusted to maximize the probability (Mitra and Acharya, 2007).

Each state transition matrix has the number "-1" as initial density. This matrix has kept changing while user's gestures are being recognized by the system.

■3.0 ANALYSYS AND OUTCOME

There are samples of results shown in following tables came from transition matrixes which are changed by occurrence of each state. The sequence of tests states is also mentioned under each of tables in Table 1 and Table 2.

Table 1 Sequence of tests states: next, next, previous, zoom-in, next

NEXT STATE / CURRENT STATE	Previous	Zoom-IN	Rotate	Zoom-OUT	NEXT
PREVIOUS	0.16	0.36	0.16	0.16	0.16
ZOOM-IN	0.16	0.16	0.16	0.16	0.36
ROTATE	0	0	0	0	0
ZOOM-OUT	0	0	0	0	0
NEXT	0.328	0.128	0.128	0.128	0.288

Table 2 Sequence of tests states: zoom-in, zoom-in, rotate, next, zoom-in, next

NEXT STATE / CURRENT STATE	Previous	Zoom-IN	Rotate	Zoom-OUT	NEXT
PREVIOUS	0	0	0	0	0
ZOOM-IN	0.1024	0.2304	0.2624	0.1024	0.3024
ROTATE	0.16	0.16	0.16	0.16	0.36
ZOOM-OUT	0	0	0	0	0
NEXT	0.16	0.36	0.16	0.16	0.16

The result of interaction with system by states sequence shown in Table1 is predicted as "Previous" and the prediction based on results shown in Table2 and states of that would be "Zoom-In" state. The change of probability is calculated as given in Equation 1. The probability of occurrence in each state can be named as state1 and 2. The relation of each state to the next one is shown in Equation 2.

$$P_{S_T} = \sum_{i=1}^{i=5} P_{S_i} = 1 \qquad \text{Equation 1}$$

$$P_{S_2} = \alpha P_{S_1}$$
$$\text{While } \alpha = 0.8\, m \qquad \text{Equation 2}$$

The probability of occurrence in each state can increase or decrease when a new gesture is recognized by the application. This system will study user behavior and predict his next action through this way.

As shown in Figure 11 (Deshpande and Karypis, 2004), when the number of experiments using this system increases the accuracy of prediction is increasing due to low change in probability of each state. The coverage ratio is decreasing with increase the number of experiments because the test set that do not have corresponding states in the higher order Markov model; thus, we have reducing their coverage (Deshpande and Karypis, 2004).

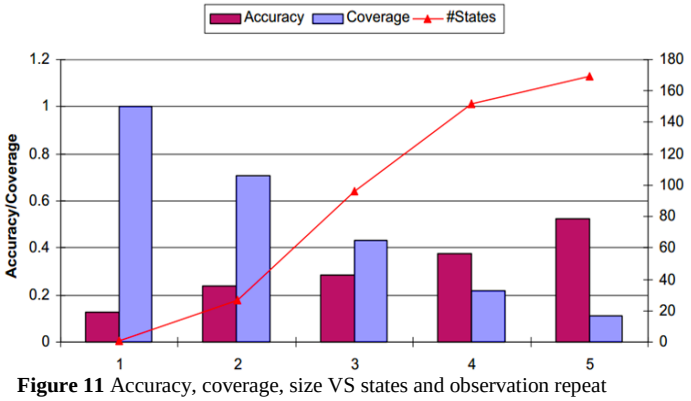


Figure 11 Accuracy, coverage, size VS states and observation repeat

4.0 CONCLUSION

The optimization of designed system in this project is the action (gesture) prediction and eases the usage of interactive systems. Although this system is designed based on using Microsoft Kinect, it is applicable to any other interactive systems that benefit other devices. It is capable to accept previous experience of a certain user and apply the same state machine. The future can be an extra engine to learn from user experiences and predict gestures not only based on HMM, but also using statistics. The high growth in the number of electronic devises shows us the importance of developing a standard for human interaction with devices and computers which is more near to natural gestures. This area has a capacity of so many researches and there are so many new methods need to make the human living environment back to its natural place while benefits technology. Prediction of such gestures based on what the user has done before can save user's energy and time.

Acknowledgement

Thanks to Multimedia University for supporting this research at Faculty of Creative Multimedia. Special thanks to Dr.Wong Chee Onn who supervised and supported me to do this research.

References

[1] Chu, C.W. and Cohen, I. 2005. Posture and Gesture Recognition Using 3D Body Shapes Decomposition. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. 69–69.
[2] Takahashi, T. and Kishino, F. 1991. Hand Gesture Coding Based on Experiments Using a Hand Gesture Interface Device. *ACM SIGCHI Bulletin*: 23: 67–74.
[3] Zimmerman, T.G., Lanier, J., Blanchard, C., Bryson, S. and Harvill, Y. 1987. A hand gesture interface device. *ACM SIGCHI Bulletin*. 18: 189–192.
[4] Nanayakkara, S., Shilkrot, R., Yeo, K. P. and Maes, P. 2013. Eyering: A Finger-Worn Input Device for Seamless Interactions with our Surroundings. *Proceedings of the 4th Augmented Human International Conference*. 13–20.
[5] Mitra, S. and Acharya, T.2007.Gesture recognition: A Survey. *IEEE Transactions on Systems, Man, and Cybernetics*. 37: 311–324.
[6] Pavlovic, V. I., Sharma, R. and Huang, T. S.1997. Visual Interpretation of Hand Gestures for Human-Computer Interaction: A Review. *IEEE*

Transactions on Pattern Analysis and Machine Intelligence. 19: 677–695.
[7] Mistry, P. and Maes, P. 2009. Sixth Sense. *TED India*.
[8] Oikonomidis, I., Kyriazis, N. and Argyros, A. 2011. Efficient Model-Based 3D Tracking of Hand Articulations Using Kinect. *British Machine Vision Conference*. 1: 1–11.
[9] Gonzalez, R.C. and Woods, R.E. 1992. Digital Image Processing. *Addison-Wesley Reading*.
[10] Webb, J. and Ashley, J.2012. Beginning Kinect Programming with the Microsoft Kinect SDK. *Apress*.
[11] Gavrila, D. M.1999.The Visual Analysis of Human Movement: A Survey. *Computer Vision and Image Understanding*. 73: 82–98.
[12] Aggarwal, J. K. and Cai, Q. 1997. Human Motion Analysis: A Review. *IEEE Nonrigid and Articulated Motion Workshop Proceedings*. 90–102.
[13] Sumar, L. and Bainbridge-Smith, A. 2011. Feasability of Fast Image Processing Using Multiple Kinect Cameras on a Portable Platform. *Department of Electrical and Computer Engineering, Univ. Canterbury, New Zealand*. 3(6).
[14] Khoshelham, K.2011. Accuracy Analysis of Kinect Depth Data. *ISPRS Workshop Laser Scanning*.1.
[15] Hai, H., Bin, L., BenXiong, H. and Yi, C. 2011. Interaction System of Treadmill Games based on Depth Maps and CAM-Shift. *IEEE 3rd International Conference on Communication Software and Networks*. 219–222.
[16] Yamato, J., Ohya, J. and Ishii, K. 1992. Recognizing Human Action in Time-Sequential Images using Hidden Markov Model. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Proceedings*. 379–385.
[17] Rabiner, L. R. 1989. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proceedings of the IEEE*. 77: 257–286.
[18] Deshpande, M. and Karypis, G. 2004. Selective Markov Models for Predicting Web Page Accesses. *ACM Transactions on Internet Technology (TOIT)*. 4: 163–184.