2011 Workshop on Digital Media and Digital Content Management

A Study on Natural Interaction for Human Body Motion using Depth Image Data

[1]Mohd Kufaisal bin Mohd Sidik, [2]Mohd Shahrizal bin Sunar, [3]Ismahafezi bin Ismail, [4]Mohd Khalid bin Mokhtar,
[5]Normal binti Mat Jusoh
UTM ViCubeLab, Department of Computer Graphics and Multimedia
Faculty of Computer Science and Information System, Universiti Teknologi Malaysia
81310, Johor, Malaysia
e-mail: [1]mohdkufaisal@gmail.com, [2]shahrizal@utm.my, [3]ismahafezi@yahoo.com, [4]khalmokh851027@yahoo.com.my,
[5]normal@utm.my

Abstract—**This paper explains a study on natural interaction (NI) in human body motion using depth image data. It involves about overview of NI and depth image. Human body motion is a non-verbal part for interaction or movement that can be used to involves real world and virtual world. Furthermore, interaction with computer or machine can be more realistic as real world and becoming more important to academic researchers, game industries, and can be adapt to other field like mechanical engineering for robotics movement and surgery purpose in medical area. Functional taxonomies will show step-by-step how human body motion were detected and created a skeleton joint. Also, we discuss about technologies behind Kinect for Xbox 360 (Kinect). Recent research in this area also included.** *(Abstract)*

Keywords-Human Computer Interaction; Natural Interaction; Natural User Interface; Human Body Motion; Depth Image Data; Depth-Aware Camera; Kinect

I. INTRODUCTION

NI [1][2][3] is a human computer interaction which targeting on some of human abilities such as body movement, touch, motion, voice, vision and using cognitive functions to interact with computer or machine. This paper focuses on NI for human body motion using depth image data.

The history of interaction and interface design is a flow and step from complex interaction to simple interaction between human and computer [4]. The word natural interaction came from Natural User Interface (NUI) that use human body interaction and voice interaction, verbal and non-verbal communication, becoming a one of Human-Computer Interaction (HCI) area. It is an evolution from Graphical User Interface (GUI). GUI is the translation from command to graphic for easier purpose for users. Before GUI era, Command Line Interface (CLI) was the starting computer interaction generation which just used codified and very strict command. Figure 1 shows the evolution of NUI.
For information, many interfaces already developed to improve this area such as Kinect [5], EyeToy [6], Microsoft Surface [7], 3D Immersive Touch [8], Dragon Naturally Speaking [9] and Perceptive Pixel [10]. Today, Kinect is one of the most popular devices used in games in NI. Kinect is used with Xbox 360 console and play with body movement. According to Microsoft's page, Kinect is designed for a revolutionary new way to play with no controller required [5]. Kinect already proved concepts of "you are the controller" [11]. In other part in this paper, we are going to tell a greatest story about succession of Kinect being one of the natural interaction devices.

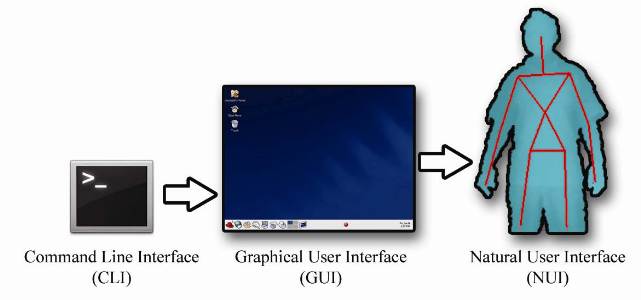

Figure 1. Evolution of NUI.

According to Calle et al. [2], HCI research has been targeting users who are not used to interacting with computers. The growing started last few years and this population of potential users brings the need for interaction systems.

In this study, we are using Kinect for the NUI and open natural interaction (OpenNI) [12] and OpenKinect [13] libraries. Both of the libraries are open-source and can be use in our testing and implementations. Figure 2 illustrates depth image data and RGB data that get from Kinect as interface and using OpenNI framework to produce the data.

Depth image data is a data that get from device call depth-aware camera like Kinect and time-of-flight (TOF) camera. This data shows like Figure 2 (left) and differentiate depth with color (bright yellow to dark yellow until black). Therefore, depth data can be measure to get distance from

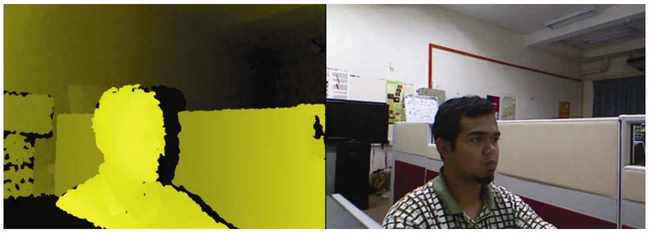

Figure 2. Depth image data (left) and RGB data (right) visualized using Kinect with OpenNI framework.

978-0-7695-4413-7/11 $26.00 © 2011 IEEE
DOI 10.1109/DMDCM.2011.26

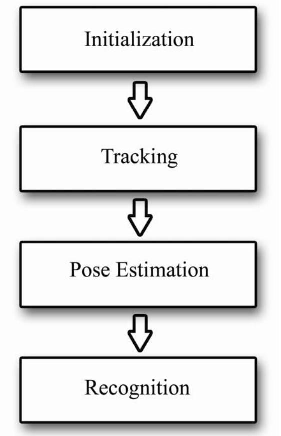

Figure 3. A general structure for systems analyzing human body motion [14].

camera to object and can be manipulate to various things like motion detection, distance measurement and others.

II. Functional Taxonomy for Human Body Motion

According to Moeslund et al. [14], functional taxonomy for human body motion has four major parts processes. The processes are initialization, tracking, pose estimation and recognition is shown in Figure 3. These processes were suggested by Moesland et al. [14]. For information, from their review article [1] based on 352 papers from year 2000 until 2006. Before review article [14] published, Moeslund et al. [15] already published survey article in year 2001. Furthermore, they compared more than 130 publications in that survey and have been cited to their review article in 2006.

System ready to process data it needs after the system being initialized. For example, appropriate model of the subject must be established. The subject must be a normal human body. These processes also can be used for disabled person but must change the techniques and methods that are used in the system. After that, the motions that get from the subject are tracked. This process did the segmenting the subject from the background and finding correspondences between segments in consecutive frames. Then, pose of the subject's body needs to be estimated. For example, control an avatar in virtual environment. The final process must analyze the pose to recognize the actions performed by the subject like walking, jumping and others movement [14][15]. The explanations above about the steps to get the human body motion were abbreviated only. All of these processes will be discusses more detail after this.

This general structure that proposed from Moeslund et al. [14] can be reduce and improve depend on techniques and methods. These processes are for vision-based human motion capture and fits perfectly with human body motion research. See Table I to get a whole view about functional taxonomies in human body motion. That table was a summarized from Moeslund et al. [15].

A. *Initialization*

Initialization process is to ensure that a system commences its operation with a correct interpretation of the current scene. Furthermore, this process is an important part of the tracking and poses estimation. A prior knowledge used in human motion capture subdivided into a number of

TABLE I. Summarized the Functional Taxonomies for Human Body Motion [15].

Functional Taxonomies for Human Body Motion	**Initialization**	Kinematic Structure Initialization	
		Shape Initialization	
		Appearance Initialization	
	Tracking	Background Subtraction	Background Representation
			Classification
			Background Updating
			Background Initialization
		Motion-based Segmentation	
		Appearance-Based Segmentation	Temporal Context-Free
			Temporal Context
		Shape-Based Segmentation	Temporal Context-Free
			Temporal Context
		Depth-Based Segmentation	
		Temporal Correspondences	Temporal Correspondences Before and After Occlusion
			Temporal Correspondences During Occlusion
	Pose Estimation	Model Free	Probabilistic Assemblies of Parts
			Example-Based Methods
		Indirect Model Use	
		Direct Model Use	Multiple View 3D Pose Estimation
			Monocular 3D Pose Estimation
			Learnt Motion Models
	Recognition	Action Hierarchies	
		Scene Interpretation	
		Holistic Recognition Approaches	Human Body-Based Recognition of Identity
			Human Body-Based Recognition
		Recognition Based on Body Parts	

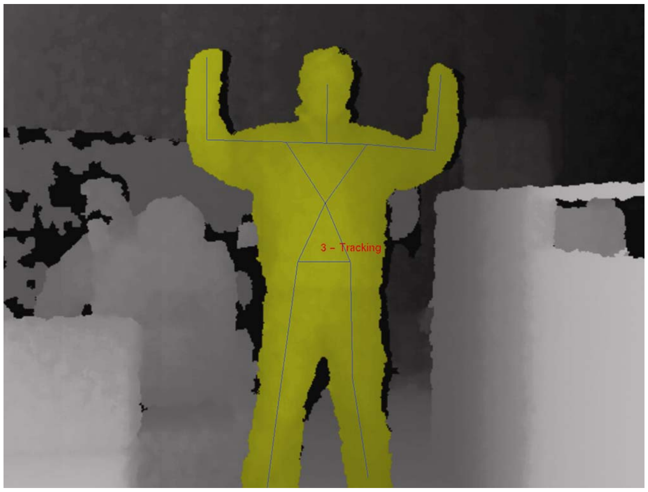

Figure 4. Calibration pose in OpenNI framework using Kinect.

sources: a kinematic structure, 3D shape, color display, pose, and type of motion [15].

In the meantime, initialization process was divided into (1) kinematic structure initialization, (2) shape initialization, and (3) appearance initialization.

Figure 4 shown a calibration pose that use in OpenNI framework. This calibration pose may change from time to time to initialize human body. This position is compulsory to do an initialization to point joints of the skeleton. If we want to make other calibration pose, we must chose an appropriate position because the sensitivity of joints detection will effects results in human body tracking.

B. *Tracking*

Tracking process is to segment and track humans in one or more frames. Since 2000, tracking algorithms that have been used have focused on surveillance applications. Methods that always applied as the first step in many tracking system are figure-ground segmentation. Fine-ground segmentation was divided into two: (1) temporal data, and (2) spatial data. These methods are the process of separating the objects of interest (human) from the rest of the image (the background) [15].

According to Moeslund et al. [15], they were categorized these methods in accordance with the type of image measurements the segmentation is based on: (1) motion, (2) appearance, (3) shape, or (4) depth image data.

Based on Tsukiyama et al. [16], they detect moving people by using moving object detection and then compare the objects to the related human library. After that, they are tracking that moving humans.

Tracking process was divided into six main parts: (1) background subtraction, (2) motion-based segmentation, (3) appearance-based segmentation, (4) shape-based segmentation, (5) depth-based segmentation, and (6) temporal correspondences. Subsequently, these main parts were broken into small parts depend on methods and functionality.

C. *Pose Estimation*

Pose Estimation is to estimate the pose of a human in one or more frames. This is the process to identify human body that is configured in a scene given. Further, this process can be an active part while doing tracking process or it will be a post processing step in tracking [15]. Referred to Moeslund et al. [15], the most difficult and ill-posed problem is the recovery of full 3D pose from single view images towards which initial steps have been made.

Hereinafter, pose estimation process was divided into three main parts: (1) model free, (2) indirect model use, and (3) direct model use. Additional, these main parts were broken into small parts depend on methods and functionality.

D. *Recognition*

Recognition is to recognize the identity of individuals as well as the actions, activities and behaviors performed by one or more humans in one or more frames [15]. The recognition aspect of human motion capture can be seen as a kind of post processing. It is relevant to include since the recognition guides the development of many motion capture systems as it is their final or long term goal. The recognition is usually carried out by classifying the captured motion as one of several types of actions. The actions are normally simple, such as walking and running, but more advanced actions such as different ballet dance steps have also been studied [14].

Recognition process was broken into five main parts: (1) action hierarchies, (2) scene interpretation, (3) holistic recognition approaches, (4) recognition based on body parts, and (5) action primitives and grammars. For information, these main parts were divided into small parts depend on methods and functionality.

III. KINECT: SUCCESFUL NATURAL INTERACTION SYSTEM

For those who always keep on track about world of technology, Kinect is one of the successful device and algorithm in natural interaction that can detect motion of human body, draw human skeleton, voice recognition, and face recognition. This product was developed by Microsoft Corporation. Kinect originally known as Project Natal before Kinect name replace it. Concept of the Kinect is "you are the controller" [11].

Figure 4 illustrates technologies behind Kinect. Kinect device have 3 main systems that can be used in Xbox 360. First, the device has a RGB camera that can take RGB data and view it like normal web camera. Second, it has Infrared (IR) camera that read the IR dots comes from projection of IR light. This camera can get depth data from environment. Lastly, Kinect also has multi-array microphone to get a human voice and able to capture even we stand far from the device. Furthermore, this device has motorized tilt to make an adjustment while detecting human body or any movement involve.

Kinect first launched at North America on November 4, 2010 [5]. Microsoft launched Kinect that can be used on Xbox 360 only not for PC. After first launched, Adafruit Industries [17] takes an initiative to get an open-source driver for Kinect by done a competition. On November 10, 2010, Héctor Martin was announced as a winner who produced a Linux driver that can access RGB camera and IR sensors. Now the open-source driver calls OpenKinect. After that, in December 2010, Primesense [18] released their open-source driver called Open Natural Interaction (OpenNI). Actually, Kinect design was based on Primesense design.

Thus, this opportunity by developed open-source driver gives us a lot of advantages to improve Kinect algorithms, techniques and methods that can be used in various fields.

IV. Human Body Joints

Human have many joints and angles. In this research, complete joints involve have 24 joints total. All the joints have their own role and functions. The joints involve: (1) head, (2) neck, (3) torso, (4) waist, (5) left collar, (6) left shoulder, (7) left elbow, (8) left wrist, (9) left hand, (10) left fingertip, (11) right collar, (12) right shoulder, (13) right elbow, (14) right wrist, (15) right hand, (16) right fingertip, (17) left hip, (18) left knee, (19) left ankle, (20) left foot, (21) right hip, (22) right knee, (23) right ankle, and (24) right foot. These joints were getting from OpenNI framework [12].

Lewis et al. [19] represented the human body model as a hierarchy of joint with skin mesh which is contains 28 DOFs for whole body and 20 DOFs for upper body. The joints of the human body that are using in system depend on functionality.

Normal joints in calibration pose just using 15 joints only without waist, right and left collar, right and left wrist, right and left fingertip, right and left ankle.

V. Recent Research

Recently, research in this area already grown up from time to time because of improvement of technologies, increases researchers involvement in this area and new techniques, methods, and algorithms have been created to facilitate other researchers to do better and exciting research soon and for the future [2].

In order to make it simple and compact, we simplified to three researches that have been done. First review is from Zhu et al. [20]. Second is from Ganapathi et al. [21]. Last review is from Parvizi et al. [22]. All of these reviews based on human body and depth image data.

A. Human Body Pose Tracking using a Bayesian Framework

Zhu et al. [20] research was used a Bayesian Framework to track human body pose using a depth image data. It was sensor [23] with frame rate at 16 frames per second (FPS). With this experiments and studies, this research has two major goals to achieve. Zhu et al. [20] list the goals as shown: (1) the proposed the Bayesian framework is able to track robustly and recover from tracking failure by integrating low-level key-point detection from depth image analysis, and (2) the proposed Bayesian framework is able to achieve a higher accuracy by taking advantage of the 3D model and point clouds [20].

TABLE II. A comparison of overall trajectory accuracy between key-point based method and Bayesian-based method [20].

Methods	X trajectory accuracy	Y trajectory accuracy	Z trajectory accuracy
Key-point based method	80 mm	84 mm	93 mm
Bayesian based method	73 mm	78 mm	87 mm

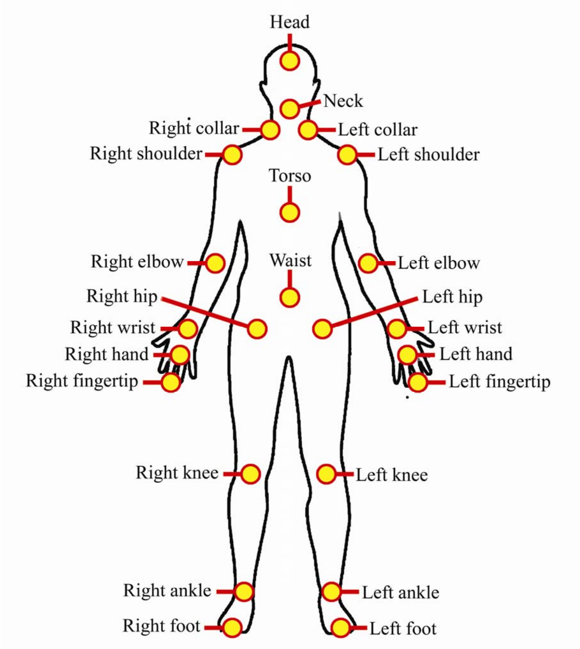

Figure 5. 24 human body joints

Current implementation from Zhu et al. [20] works well for body when twists up to 40 degree rotation on either side or in front of facing posture. Large twists and severe interaction between upper and lower body limbs remain as a challenge in the current implementation.

For the results, Zhu et al. [20] summarized the experiments that have been made and compared its performance with other two methods, Iterative Closest Point (ICP) based method [24] and key-point based method as in Table II. A Bayesian-based method shows better result than other two other methods. By using the Bayesian-based method, Zhu et al. [20] proved that this method can tracking

TABLE III. Comparison between various human pose tracking approaches [20].

Methods	Tracking through occlusion	Error-recovery	Tracking with missing key-points	Integration with other information	Speed
ICP based method	No	No	Yes	No	5~9 Hz
Key-point based method	Yes	Yes	No	No	3~6 Hz
Bayesian-based method	Yes	Yes	Yes	Yes	0.1 Hz

through occlusion, error-recovery, tracking with missing key-points, integration with other information and speed is better than ICP based method and key-point based method. Furthermore, the Bayesian-based method proved that this method can achieve a better accuracy for joint trajectories rather than use key-point base methods. Table III shows the results from two methods, key-point based method and the Bayesian based method. It is roughly shows that accuracy using the Bayesian based method is more accurate because the value of distance less (6~7 mm) less than other method.

B. Real-Time Motion Capture using Depth Camera

Firstly, Ganapathi et al. [21] actually predicts the future about motion capture technology become convenient, cheap, and applicable in natural interaction environment. Nowadays, technologies become cheaper, and more reliable like Kinect. Even though before this the only solution of human motion capture was marker-based, but now, marker less motion capture area already become more popular systems [25]. They propose in their research paper a probabilistic filtering framework that employs a highly accurate generative model which is achievable in this setting using an efficient GPU implementation with a discriminative model.

They specifically developed their algorithm for fast video frame rates. Actually, natural communication requires a low latency action-reaction cycle. In their presented system, estimation for joint angles of a 48 degree-of-freedom (DOF) human model by requiring four to ten frame per second (FPS) [21].

Based on Ganapathi et al. [21] were highlighted two contributions inside their research paper. First, they contribute a novel algorithm for combining local hill climbing and discriminative part detections. Second contribution is a definition of a smooth likelihood function and a means of implementing it on readily available graphics hardware (GPUs) efficiently in order to obtain near real-time performance [21].

Ganapathi et al. [21] list the goals as shown: (1) our proposed system is able to estimate the pose and configuration of a human over time using only a stream of depth images, (2) proposing candidates using EP on detected body parts significantly improves performance over just doing local hill-climbing, (3) the smoothed energy function outperforms the typically used pixel-wise energy function, and (4) the system runs close to real-time.

In their experiment, they use depth image data that get from Swissranger SR-4000 TOF camera [26]. The results from their experiment after test three different algorithms, expectation propagation (EP), hill climbing (HC) and combined approach (HC+EP). EP approach gave worse results than the other two approaches in their average pose error graph. HC+EP approach performs best or equally well with just use HC only. In terms of efficiency, HC+EP ran at 4-6 FPS while HC approach ran at 6 FPS. So, it is look like HC more efficient than combined algorithm but actually both algorithms are close.

C. 3D Head Tracking in Real-Time

This research paper writes by Parvizi et al. [22] with title Real-Time 3D Head Tracking Based on Time-of-Flight Depth Sensor conducted experiments based on the proposed method, contour analysis based on depth image data, and used time-of-flight (TOF) sensor [23].

Parvizi et al. [22] also claimed their research paper introduced a novel head detection technique for head tracking using TOF sensor to generate the input range images and incorporates contour analysis. The algorithm detects moving regions by subtracting two consecutive frames from each other and highlighting the difference [21].

Actually, the results of this research proved that the system is robust against drastic illumination variations since range data is less dependent to ambient light than intensity images [22]. Parvizi et al. [22] also demonstrated that by incorporating TOF sensor, we can reduce the computational cost and the complexity of the system because we no longer needed to model the background.

With this explanation, by using depth image data that have been provide by depth-aware camera like TOF or Kinect will produce the better results, much easier and reduce our computational cost. If we use normal camera, it will gain our computational process and must make an optimization to reduce the memory resources and processes time.

VI. Conclusions and Discussions

We already presented the study in natural interaction area where focus on human body motion and depth image data for the input while in the tracking. The latest technology about Kinect has been mention in this paper. Use of Kinect not only narrow on Xbox 360 console only but can be use at our own computer. With the professional skills, Héctor Martin, we can use the opportunity from him to evolve this area. Nowadays, natural interactions becoming more interesting area and have a bright future to bring our world to the biggest steps in an interaction with virtual environment [2].

Utilization of depth image data will reduce computational cost because we actually eliminate one step of removing background from object. Thus, the object can be use directly from source. This also can reduce processing time.

Table IV shows a summarized data from recent researches that we have been done. It is divided into five major header, year, first author, camera type, body parts and algorithm that were used in their research. All of these researches use same camera brand [23][26] because during

TABLE IV. Summarized from recent researches that were reviews.

Year	First author	Camera type	Body parts	Algorithm
2010	Zhu et al. [20]	SwissRanger SR-3000 [23]	Whole body	Bayesian Framework
2010	Ganapathi et al. [21]	SwissRanger SR-4000 [26]	Whole body	HC+EP
2007	Parvizi et al. [22]	SwissRanger SR-3000 [23]	Head	Contour analysis

that time, technologies from Primesense [18] and Kinect [5] not so popular at that day or not born yet to HCI technologies arena. Kinect much cheaper than SwissRanger and same results will generate.

Improvement in human body motion needs to be done from time to time because it can be apply in other area of knowledge. Potential application from this research will come out with the alternatives motion capture. This motion capture will pledge to be a low-cost motion capture and can be used in many ways such as surveillance and 3D animation.

ACKNOWLEDGMENT

This research was supported by UTM VicubeLab at Department of Computer Graphics and Multimedia, Faculty of Computer Science and Information System, Universiti Teknologi Malaysia. Special thanks to Ministry of Higher Education (MOHE) Vot. 78658 grant for providing financial support of this research.

REFERENCES

[1] Valli, A. (2005). "Notes on Natural Interaction."

[2] Calle, J., P. Martínez, et al. (2009). "Towards the Achievement of Natural Interaction." Engineering the User Interface, Springer London**:** 1-19-19.

[3] Del Bimbo, A. (2008). "Special issue on natural interaction." Multimedia Tools and Applications 38(3): 293-294-294.

[4] Valli, A. (2008). "The design of natural interaction." Multimedia Tools Appl. 38(3): 295-305.

[5] ."Kinect for Xbox 360 (Kinect)." from http://www.xbox.com/en-US/Kinect.

[6] ."EyeToy." from http://www.eyetoy.com.

[7] ."Microsoft Surface." from http://www.microsoft.com/surface.

[8] ."3D Immersive Touch." from http://www.immersivetouch.com.

[9] ."Dragon Naturally Speaking." from http://www.nuance.com/for-individuals/by-product/dragon-for-pc/index.htm.

[10] ."Perceptive Pixel." from http://www.perceptivepixel.com.

[11] ."Kinect Ads: 'You Are the Controller'." from http://www.microsoft.com/presspass/features/2010/oct10/10-21kinectads.mspx.

[12] ."Open Natural Interaction (OpenNI)." from http://www.openni.org.

[13] ."OpenKinect." from http://www.openkinect.org.

[14] Moeslund, T. B. and E. Granum (2001). "A survey of computer vision-based human motion capture." Comput. Vis. Image Underst. 81(3): 231-268.

[15] Moeslund, T. B., A. Hilton, et al. (2006). "A survey of advances in vision-based human motion capture and analysis." Comput. Vis. Image Underst. 104(2): 90-126.

[16] Tsukiyama, T. and Y. Shirai (1985). "Detection of the movements of persons from a sparse sequence of TV images." Pattern Recognition 18(3-4): 207-213.

[17] ."Adafruit Industries." from http://adafruit.com/.

[18] ."Primesense." from http://www.primesense.com/.

[19] Lewis, J. P., M. Cordner, et al. (2000). "Pose space deformation: a unified approach to shape interpolation and skeleton-driven deformation." Proceedings of the 27th annual conference on Computer graphics and interactive techniques, ACM Press/Addison-Wesley Publishing Co.: 165-172.

[20] Zhu, Y. and K. Fujimura (2010). "A Bayesian Framework for Human Body Pose Tracking from Depth Image Sequences." Sensors 10(5): 5280-5293.

[21] Ganapathi, V., C. Plagemann, et al. (2010). "Real time motion capture using a single time-of-flight camera." Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on.

[22] Parvizi, E. and Q. M. J. Wu (2007). "Real-Time 3D Head Tracking Based on Time-of-Flight Depth Sensor." Proceedings of the 19th IEEE International Conference on Tools with Artificial Intelligence - Volume 01, IEEE Computer Society: 517-521.

[23] ."SwissRanger SR-3000." from http://www.swissranger.ch.

[24] Besl, P. J. and N. D. McKay (1992). "A Method for Registration of 3-D Shapes." IEEE Trans. Pattern Anal. Mach. Intell. 14(2): 239-256.

[25] Poppe, R. "Vision-based human motion analysis: An overview." Computer Vision and Image Understanding 108(1-2): 4-18.

[26] ."SwissRanger SR-4000." from http://www.swissranger.ch.

[27] Anguelov, D., P. Srinivasan, et al. (2005). "SCAPE: shape completion and animation of people." ACM Trans. Graph. 24(3): 408-416.

[28] Ismail, I., Sunar, M. S., Mohd Sidik, M. K. and Mohktar, M. K. "A Review of Dynamic Motion Control Considering Physics for Real Time Animation Character" Proc. Workshop on Digital Media and Digital Content Management (DMDCM 2011), in press.

[29] Mokhtar, M. K., Sunar, M. S., Mohd Sidik, M. K., Ismail, I. "Hierarchical Occlusion Queries on Driving Simulator" Proc. Workshop on Digital Media and Digital Content Management (DMDCM 2011), in press.