Context-Aware 3D Gesture Interaction Based on Multiple Kinects

Maurizio Caon[1,2], Yong Yue[1]
[1]Faculty of Creative Arts, Technologies & Science
University of Bedfordshire
Luton, United Kingdom
e-mail: {maurizio.caon, yong.yue}@beds.ac.uk

Julien Tscherrig[2], Elena Mugellini[2], Omar Abou Khaled[2]
[2]Department of Informatics
University of Applied Sciences of Western Switzerland, Fribourg
e-mail: {julien.tscherrig, elena.mugellini, omar.aboukhaled}@hefr.ch

***Abstract*—This paper presents a novel context-aware system for deictic gestures interaction with smart environments. The system tracks multiple users; moreover, it recognizes inhabitants' postures and gestures in real-time. This information, enriched with smart objects coordinates, is reconstructed in a 3D model to allow the recognition process. Finally the system executes the programmed tasks to support the users' activity. Two Microsoft Kinect depth cameras have been used to acquire the data and a framework for the communication with the smart objects has been adopted. A first prototype has been developed and an evaluation test with 13 users has been conducted in order to assess the usability of the system. Results show that this interaction experience has been really appreciated by the users.**

Keywords-Ambient Intelligence; Smart Environment; Gesture Interaction; Posture Recognition; Depth Cameras

I. INTRODUCTION

Norman's invisible computer [1] and Weiser's ubiquitous computer [2] theories led to the conception of Ambient Intelligence (AmI). This multidisciplinary paradigm is intended to create new smart infrastructures that seamlessly integrate intelligent services [3]. This new conception of computing is thought to be invisible but always active in the background to grant all the services that are supposed to be necessary to the user. This novel anthropomorphic human-machine model of interaction permits the user to move into the foreground in complete control of the smart, augmented environment which interprets actions to support and enhance the abilities of its occupants in executing tasks [4]. Such a system has awareness about the user's current activity, situation and intention before the activity is actually completed to provide appropriate support. A technology that allows reading the human mind does not exist yet, and capturing actor's intention implicitly is a very hard challenge [5]. For this reason, a smart environment cannot limit its decisions to the elaboration of context information, but it has to allow the inhabitants to interact with the environment in order to explicit their intentions and goals. According to the current trends, the design of an interactive environment interface aims especially to solve important issues related to the usability and the adaptiveness of such interfaces. In order to create an end-user friendly interface, many researchers are focused on natural ways of interaction as speech and gestures [6]. In particular, deictic gestures play an important role since they are intuitive and commonly used by humans to reference objects and devices by pointing at them [7]. Therefore, deictic gestures are really significant for human-environment interaction. On the other hand, correct deictic gestures interpretation by the system depends on the user context (e.g., position and orientation in the 3D space) and this involves the need of a situational awareness about the smart objects placed in this interactive space.

A real-time context-based system for deictic gestures interaction with smart environments is presented in this paper. This system grants human actors to interact with a smart environment pointing at the smart objects. The context information comes from the data related to the states and positions of the smart objects, and to the tracked inhabitants' postures. All this information is modeled in a 3D virtual space. The sensors that are used to acquire the data are two Microsoft Kinect depth cameras. Hence the user does not need to wear any special device.

The 3D camera technology is positioned to become ubiquitous; in fact the Microsoft Kinect is a really cheap off-the-shelf device which can provide quite accurate depth information (11 bit data for 2,048 levels of sensitivity) at a good frame-rate (30 Hz). The research community has already manifested strong interest in this new device, also for applications that go far beyond simple video-gaming [8][9][10].

The rest of the paper is organized as following. In Section 2, the related work is presented and the scenario is described in Section 3. The tests and the considerations related to the use of multiple Kinects are discussed in Section 4; the architecture of the system is described in Section 5. Section 6 presents the tests that have been made in order to evaluate the system. Section 7 is dedicated to the conclusion reporting also the future work.

II. RELATED WORK

Interaction between human beings and smart environments is a real demanding research area [6]. Gestures are really significant for this kind of applications [11].

Many research works are focused on deictic gestures and often they prefer cameras as sensors to capture the inhabitants' movement information, as in [12] and [13].

Copyright (c) IARIA, 2011. ISBN: 978-1-61208-170-0

Information extraction from 2D video streams involves many limitations because pointing in a real room needs 3 dimensions for a complete representation. In fact, the authors of [14] adopted stereo-cameras to extract depth information from disparity map.

Postures recognition is another research domain that captured researchers' interest [15]. Most of the works adopted video-cameras as sensing device [16][17][18]. The elaboration of data deriving from a single 2D video flow involves many problems, e.g., occlusion and cluttered background, and puts many limits in recognizing human postures. Indeed, Chu and Cohen adopted a system based on four synchronous cameras to recognize postures and gestures [19]. Another solution to add important spatial information to recognize postures consists in using a depth camera, as in [20].

Depth cameras can really improve system performances for the interaction in a smart environment, as Wilson and Benko demonstrated in [21]. On the other hand, this kind of systems allows the interaction only with selected surfaces.

In this paper, a novel system that recognizes 3D deictic gestures and human postures using multiple depth cameras is presented. It reconstructs context information elaborating spatial coordinates of the smart objects (that are previously inserted in the system) and of the inhabitants (that are constantly tracked in the 3D space), combining them with the deictic gestures and postures data to improve the interaction experience.

III. SCENARIO

Youngblood et al. defined a smart environment as one that is able to acquire and apply knowledge about the environment and its inhabitants in order to improve their experience in that environment [22]. Designing a smart room that can achieve this goal involves the context awareness and the possibility of interaction with the people. Gestures are a natural way of interaction for humans and integrating these commands with information coming from the situation can make the environment to support user's tasks. A smart room that can achieve this goal has to recognize the inhabitants' activity, it has to understand the direct commands ordained by the users and it must integrate many smart objects to communicate with. Therefore, the smart environment detects the human posture and the smart objects state; indeed it can create context information. Commands, acquired from the users present in the environment, are interpreted referring to the previously modeled context. Our context information comes from the data related to the states and positions of the smart objects, and to the tracked inhabitants' postures.

The target scenario deals with a smart living room which recognizes deictic gestures and human postures, tracks the inhabitants and allows them to interact with smart objects that are present in the interactive space, see Figure 1. The novel approach of this system assigns different meanings to the pointing gesture according to the user's posture. If the user is sitting on the couch in front of the TV and points at the media center, then the TV is turned on. If the user is standing in the center of the room and he points at the media center, then the radio is turned on (the environment is set to interpret this situation as the user would like to listen to music). The spatial coordinates and the postures are crucial for the interpretation of the meaning of the gestures performed by the user, but the state of the smart objects is important as well. If the human points at the lamp, which is turned off, then the system turns it on; on the contrary if the lamp is already on, the system turns it off. Another application deriving from the posture recognition part is aimed to emergency purposes. When the system detects a person lying on the ground for a period superior to a guard time that has been previously set, the system calls the rescues. On the other hand if the human is lying on the sofa, the smart environment closes the blinds and turns off the lights to permit to the user to rest comfortably.

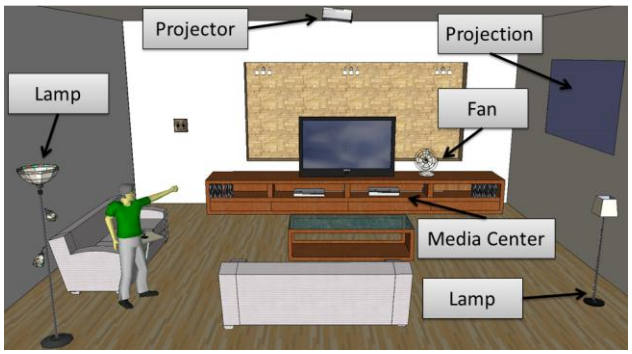

Figure 1. Scenario.

IV. USING MULTIPLE KINECTS

Robust interactive human body tracking has many applications, in particular it can be important in human-computer interaction; depth cameras can simplify reaching this task and the Microsoft Kinect represents the first down-market device that can allow capturing 3D general body motions and shapes at interactive rates [20]. However, using multiple Kinects involves interference between the infrared laser patterns that are at the base of the functioning of this device. Each Kinect projects its own infrared pattern for the calculation of the depth information and interferences can degrade the information quality creating black spots on the 3D image. In order to assess if the interferences change significantly referring to the number of active Kinects and their positions, 5 different configurations have been tested. Figure 2 represents the camera configurations that have been tested. The Kinects have been positioned in A, B and C. The colored triangles represent the field of view of the cameras from the A, B and C positions. The striped areas of the triangles represent the interactive areas, or rather, the areas where people can be easily tracked. This area begins at a distance of 0.8 m from the Kinect and arrives to 3.5 m. A person was present in the test scenario and he was positioned on the white circle in the center of the figure. The A, B, and C positions are at the same distance from the person, in order that the points of the patterns projected from the infrared lasers have same brightness and dimensions on the person. The optical axis of the Kinect positioned in A intersects the optical axis of the Kinect

Copyright (c) IARIA, 2011. ISBN: 978-1-61208-170-0

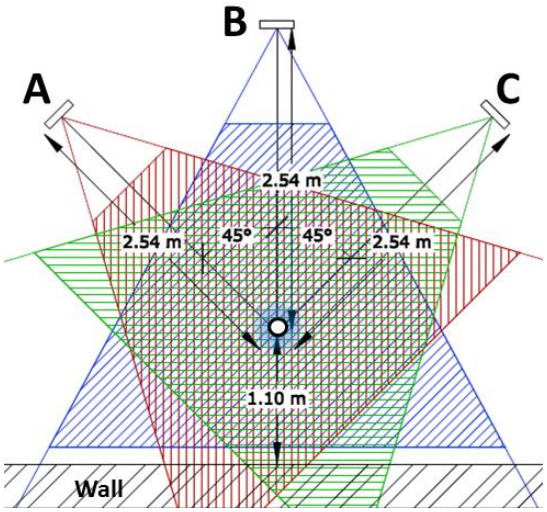

Figure 2. Representation of the interference test configurations.

positioned in B forming an angle of 45°. The optical axis of the Kinect positioned in A intersects perpendicularly the optical axis of the Kinect positioned in C. In configuration 1 there was only one active Kinect and it was positioned in A. In configuration 2 there were two active Kinects and they were both positioned in A. In configuration 3 there were two active Kinects, one was positioned in A and the other one in B. In configuration 4 there were two active Kinects, one was positioned in A and the other one in C. In configuration 5 there were three active Kinects, one was in A, one in B and one in C. In Figure 3, the captures for every configuration taken from the Kinect that has always been in A are reported. The black pixels in the captures represent the pixels without depth information.

To quantify the interference effect, the number of pixel without depth information has been calculated. Since the pixels without depth information change during time also on a static scene, then this number has been calculated making an average on 1000 frames for every configuration. The depth sensor of the Kinect captures 640x480 pixel frames; therefore, every frame has got 307200 pixels with depth information. For the configuration 1, an average of 5325 pixels without depth information has been calculated (with standard deviation of 165 pixels); for the configuration 2 the average was of 14018 pixels and the standard deviation was of 404 pixels; for the configuration 3 the average was of 12502 pixels and the standard deviation was 319 pixels; for the configuration 4 the average was of 13000 pixels and the standard deviation was of 295 pixels; for configuration 5 the average was of 21813 pixels and the standard deviation was of 432 pixels. After these tests, we verified that the interference caused by two Kinects is not significant for the skeleton tracking and it remains almost constant regardless the relative position of the two cameras. However, using two Kinects in configuration 4 permits capturing the tracked users' movements from very different perspectives. This configuration permits to capture a very big portion of the users' bodies avoiding in many cases the occlusion of some

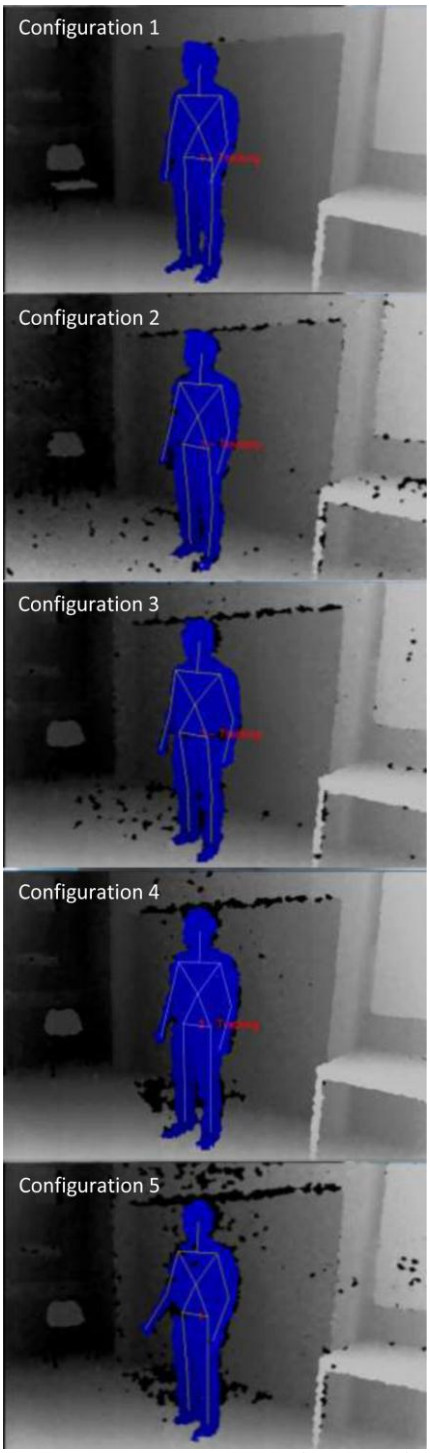

Figure 3. Scene captures during the interference test for every configuration.

limbs. In Figure 4 a), an example is shown; on the upper part the Kinect can capture the left side of the user and cannot see the legs; in the lower part of this figure there is the view from the other Kinect present in the system that can capture the information about user's legs, but it cannot track his left arm. The system combines the data coming

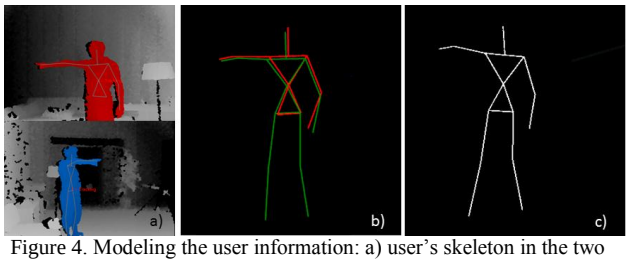

Figure 4. Modeling the user information: a) user's skeleton in the two Kinect views; b) 3D model of the user's skeletons captured by the two Kinects; c) 3D fusion of the user's two skeletons in one skeleton

from the two Kinects to reconstruct the whole user's skeleton as explained in the next section.

Using three Kinects in configuration 5 doubles the number of the pixels without depth information generated by the interference but it does not add significant information for the user's skeleton reconstruction. More tests with three and four Kinects surrounding the users will be executed soon.

According to the results of these tests, in our system we decided to use two Kinect cameras positioned as in configuration 4. The developed system is presented in the next section.

V. SYSTEM ARCHITECTURE

The developed system has been designed to be modular. It consists of two acquisition modules (one for each Kinect camera) and a central module.

The acquisition modules are based on the OpenNI libraries [23] and can track the people present in the vision area. Each module constructs a skeleton model of each tracked user that will be represented in a specific XML structure. These data will be sent to a central module for the 3D modeling. These modules communicate with XML messages using the UDP protocol. Every XML message contains the information about the coordinates of every joint of every tracked person (that is uniquely identified) and the ID number of the Kinect camera that sent these data. Moreover, in the XML message there is also the destination IP address since this system has been designed to communicate with other machines in order to make possible a future distributed version to spread the global interactive area.

The central module makes the fusion of the data concerning the same tracked user to create a 3D skeleton, as shown in Figure 4. This 3D skeleton is calculated with coordinates of the different joints and it is placed in the 3D model of the environment. The fusion algorithm calculates the difference between the coordinates of every joint, and then it makes an average of these coordinates weighting the information of the more reliable data. In fact, the system assigns a higher weight to the coordinates that come from the Kinect capturing the highest number of joints information. Depending on the user's position and posture, a Kinect can see a bigger or smaller part of the user's body. When a Kinect cannot track a specific part of the body to calculate the joints coordinates, it does not send any information to the central module about that joint. For this reason, the algorithm for the fusion assigns a higher weight to the Kinect that can detect more joints. When the coordinates of a specific joint are provided from only one Kinect because the other one cannot track this body part, the algorithm uses directly the unique received data. When both Kinects cannot provide the coordinates of a specific joint, then the system ignores that joint waiting for new data.

The data related to the coordinates of the smart objects present in the interactive area must be inserted in the 3D model of the environment using a dedicated interface of the central module. The advantage of this data representation consists in the possibility of setting spatial constrains in 3 dimensions for the interaction with the smart objects.

A. Calibration

The cameras calibration is crucial to reconstruct a 3D model using simultaneously multiple depth cameras. The coordinates of the joints coming from the two cameras must be represented in the 3D reconstruction of the environment. In order to find the right relative coordinates of the points captured from the two different points of view, a transformation matrix must be determined. The transformation matrix in 3D is a special 4x4 matrix and is based on quaternions [24]. This transformation matrix provides a rotation around the x, y and z axis. The calibration phase aims to calculate the 16 values of this matrix. To solve this matrix it is necessary to obtain the coordinates of 4 fixed points from the two cameras. The resolution of quaternion matrix aims to resolve the transformation point from a relative coordinated system to the main coordinated system.

Considering four common points on the two different Kinect cameras with known exact coordinates, then it is theoretically possible to calculate the transformation matrix.

The calculation of the transformation matrix based on quaternion aims to resolve 16 equations with 16 unknowns. All the calibration process is effectuated from the central module.

The cameras calibration phase must be effectuated during the set-up of the system. The calibration permits to position the two Kinects in every desired configuration, the only constrain is that they must capture the same scene. In fact, the two Kinects must see at least four common points to accomplish the calibration. This system has been programmed to register up to 10 common points to calculate several transformation matrixes. Afterwards, the system computes for each transformation matrix the average delta between the main coordinates and the transformed coordinates of all the captured points; therefore, the transformation matrix with minimum average delta is chosen. The system can utilize the calculated transformation matrix as long as the Kinects remain in the same positions.

B. Gesture and Postures Recognition

The 3D model of the environment includes the users' skeletons and the smart objects. The system recognizes the users' postures and pointing gestures from the coordinates of the joints in real-time. This recognition process is based on simple conditions referred to some values of the joints and

Copyright (c) IARIA, 2011. ISBN: 978-1-61208-170-0

the relative distances between them. The postures that are recognized by the system are two: standing and sitting. This information is completed with the relative positions of the objects in the 3D model of the environment. When the joints of the arm assume specific values, the system projects a prolongation of the arm and calculates if it intersects the active area attributed to a specific object. Our system has been integrated in the NAIF framework [25]. NAIF framework handles the creation and management of a smart environment. It manages the set-up and the communication between smart objects and devices present in the environment. Thanks to NAIF, our system can check the current object state and generate the suitable command, e.g., if the lamp state is off then it sends the appropriate command to turn it on.

VI. System Evaluation

In order to have a feedback about the system usability and to understand the limitations of our prototype, we performed an evaluation test composed of two phases. The subjects of this test are 13 users (9 men and 4 women) with different backgrounds and origins, and with age between 19 and 28 years.

A. First Phase

The subjects have been conducted to the smart living room where a simple scenario has been prepared. One user at a time has been asked to enter in the room and to interact with the system (see Figure 5). After the skeleton tracking initialization stage (the user has to remain in a pose for few seconds in front of each Kinect device), the user had to point at a lamp to turn it on; afterwards the user had to point at the media center to turn on the radio and later he had to do it again to turn it off. Afterwards, he had to sit down on the couch and to point at the media center to turn on the TV. The system never failed the gesture or the posture recognition during the test. Once finished the interaction session in the smart living room, every subject evaluated the experience through a System Usability Scale (SUS) [26] questionnaire rating the system features according to a 5-point Likert scale. The statements covered a variety of aspects of system usability, such as the need for support, training, complexity, efficiency (how much effort is necessary in achieving those objectives) and experience satisfaction.

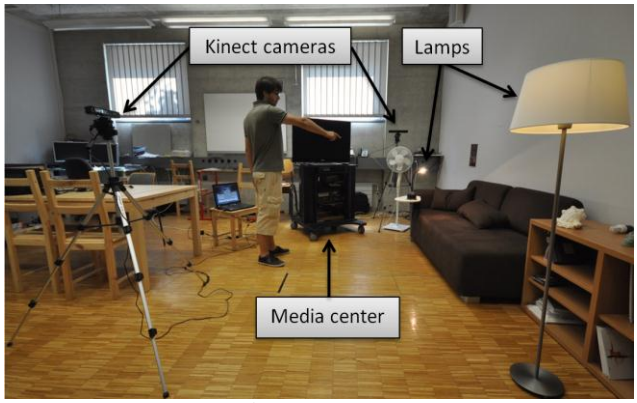

Figure 5. One of the users testing the system.

The users' evaluations assessed the system usability as excellent with an average SUS score of 90.6 points and a standard variation of 5 points.

B. Second Phase

This phase consisted in an interview where the users have been asked to express their impressions and suggestions. Most of them said that the skeleton tracking initialization stage could be really annoying for an everyday interaction in a real smart room. The subjects have been asked to say if they have missed the voice interaction modality in this test scenario and everybody answered negatively, moreover they expressed their appreciation about this interaction modality through deictic gestures. Some of the users remarked that they would like also other gestures to go beyond the turning on or off the household appliances, e.g., they would like to interact with the media center to change TV program or the volume.

Another limitation that came from our analysis is the pre-determined set of tasks that the system executes referring to the users' gestures and postures. In fact, a system that automatically learns user's habits could be preferable to a programmed one. Therefore, in order to make this system more human-centered, the integration of learning algorithms has been thought in order that the system can learn users' habits.

VII. Conclusion and Future Work

In this paper, a real-time context-based system for deictic gestures interaction with smart environments has been presented. The system tracks multiple users and reconstructs situational information collecting data about the people's postures and coordinates. Moreover, this software realizes a 3D model of the environment where only the tracked users and the smart objects are present. The users' skeletons are modeled referring to the joints coordinates captured by two calibrated Microsoft Kinect cameras; the smart objects coordinates have been inserted previously in the system and their current states are provided by NAIF framework. The 3D model of this information makes the postures and deictic gestures recognition easy. The context awareness makes possible to interpret the pointing gesture referring to the posture and coordinates of the user, giving different meanings to the same gesture in order to execute different tasks. In the current prototype the tracked people can point at two lamps and at the media center. Pointing at the lamps turns them on or off (it depends from the previous state). Pointing at the media center turns on or off the radio if the user is standing, otherwise turns on or off the TV if he is sitting on the couch. The usability tests assessed that this interaction modality with the smart environment is really intuitive; indeed the users do not need training to interact with the smart objects and they affirmed that they had a really pleasant experience. Future works are already planned in order to shorten, or if possible, eliminate the skeleton tracking initialization stage (resulted annoying for the users

Copyright (c) IARIA, 2011. ISBN: 978-1-61208-170-0

during the evaluation tests) and to add the learning algorithms for a more human-centered system that learns users' habits. Afterwards, more gestures for an augmented environment control will be implemented. Adding more gestures will increase the recognition complexity and the precision of the Microsoft Kinect could become critical, for this reason a comparison with a ground truth will be conducted. Finally, the system will be tested with multiple users interacting with the environment at the same time.

REFERENCES

[1] D. A. Norman, The Invisible Computer, Cambridge, MA: MIT Press, 1999.

[2] M. Weiser, The Computer for the 21st Century, Scientific American, September 1991.

[3] N. Shadbolt, "Ambient intelligence," IEEE Trans. Intell. Transp. Syst., vol. 18, no. 4, pp. 2–3, Jul.–Aug. 2003.

[4] P. Remagnino and G.L. Foresti, "Ambient Intelligence: A New Multidisciplinary Paradigm," IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans, vol. 35, Jan. 2005, pp. 1-6.

[5] D. Surie, T. Pederson, F. Lagriffoul, and L. E. Janlert, "Activity Recognition using an ' Egocentric ' Perspective of Everyday Objects," Time, 2007, pp. 1-16.

[6] D. Cook and S. Das, "How smart are our environments? An updated look at the state of the art," Pervasive and Mobile Computing, vol. 3, Mar. 2007, pp. 53-73.

[7] M. Karam, "A taxonomy of gestures in human computer interactions," 2005, pp. 1-45.

[8] E. Suma, D. Krum, B. Lange, S. Rizzo, and M. Bolas, "Faast: The flexible action and articulated skeleton toolkit," IEEE Virtual Reality, 2011, pp. 247-248.

[9] C. Schönauer, T. Pintaric, and H. Kaufmann, "Full body interaction for serious games in motor rehabilitation," Proc. 2nd Augmented Human International Conference, ACM, 2011, p. 4.

[10] A. DeVincenzi, L. Yao, H. Ishii, and R. Raskar, "Kinected conference: augmenting video imaging with calibrated depth and audio," Proc. ACM 2011 conference on Computer supported cooperative work, ACM, 2011, p. 621–624.

[11] A.M. Rahman, M.A. Hossain, J. Parra, and A. El Saddik, "Motion-path based gesture interaction with smart home services," Proc. of the seventeen ACM international conference on Multimedia (MM '09), 2009, p. 761.

[12] R. Kehl and L. Van Gool, "Real-time pointing gesture recognition for an immersive environment," Proc. Sixth IEEE International Conference on Automatic Face and Gesture Recognition, 2004, pp. 577-582.

[13] A. D. Wilson and a F. Bobick, "Recognition and interpretation of parametric gesture," Proc. Sixth IEEE International Conference on Computer Vision, 1998, pp. 329-336.

[14] E. Seemann and R. Stiefelhagen, "3D-tracking of head and hands for pointing gesture recognition in a human-robot interaction scenario," Proc. Sixth IEEE International Conference on Automatic Face and Gesture Recognition, 2004, pp. 565-570.

[15] E. Farella, A. Pieracci, L. Benini, and A. Acquaviva, "A Wireless Body Area Sensor Network for Posture Detection," Proc. 11th IEEE Symposium on Computers and Communications (ISCC'06), 2006, pp. 454-459.

[16] S. Mu and I. Lii, "A multiscale morphological method for human posture recognition," Proc. Third IEEE International Conference on Automatic Face and Gesture Recognition, 1998, pp. 56-61.

[17] N. Zouba, B. Boulay, F. Bremond, and M. Thonnat, "Monitoring activities of daily living (adls) of elderly based on 3d key human postures," Cognitive Vision, 2008, p. 37–50.

[18] L.B. Ozer and W. Wolf, "Real-time posture and activity recognition," Proc. Workshop on Motion and Video Computing, 2002, pp. 133-138.

[19] C.W. Chu and I. Cohen, "Posture and gesture recognition using 3D body shapes decomposition," Human-Computer Interaction, 2005.

[20] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake, "Real-time human pose recognition in parts from single depth images," In CVPR, 2011.

[21] A.D. Wilson and H. Benko, "Combining multiple depth cameras and projectors for interactions on, above and between surfaces," New York, New York, USA: ACM Press, 2010.

[22] G.M. Youngblood, D.J. Cook, L.B. Holder, E.O. Heierman, "Automation intelligence for the smart environment," Proc. International Joint Conference on Artificial Intelligence, 2005.

[23] http://www.openni.org/ 07.08.2011

[24] B.K.P. Horn, "Closed-form Solution of Absolute Orientation using Unit Quaternions," Journal of the Optical Society of America, Series A, 4, 4, 1987, pp. 629-642.

[25] D. Perroud, F. Barras, S. Pierroz, E. Mugellini, and O. Abou Khaled, "Framework for development of a smart environment, Conception and Use of the NAIF Framework," Proc. 11th International Conference on New Technologies of Distributed Systems (NOTERE), Paris, France, 2011.

[26] J. Brooke, "SUS-A quick and dirty usability scale," Usability evaluation in industry, 1996, pp. 189–194.

Copyright (c) IARIA, 2011. ISBN: 978-1-61208-170-0