Kinect Based Calling Gesture Recognition for Taking Order Service of Elderly Care Robot

Xinshuang Zhao, Ahmed M. Naguib, and Sukhan Lee*, *Fellow, IEEE*

***Abstract*— This paper proposes a Kinect-based calling gesture recognition scenario for taking order service of an elderly care robot. The proposed scenarios are designed mainly for helping non expert users like elderly to call service robot for their service request. In order to facilitate elderly service, natural calling gestures are designed to interact with the robot. Our challenge here is how to make the natural calling gesture recognition work in a cluttered and randomly moving objects. In this approach, there are two modes of our calling gesture recognition: Skeleton based gesture recognition and Octree based gesture recognition. Individual people is segmented out from 3D point cloud acquired by Microsoft Kinect, skeleton is generated for each segment, face detection is applied to identify whether the segment is human or not, specific natural calling gestures are designed based on skeleton joints. For the case that user is sitting on a chair or sofa, correct skeleton cannot be generated, Octree based gesture recognition procedure is used to recognize the gesture, in which human segments with head and hand are identified by face detection as well as specific geometrical constrains and skin color evidence. The proposed method has been implemented and tested on "HomeMate", a service robot developed for elderly care. The performance and results are given.**

I. Introduction

Elderly care robot has become a trend since some of the developed countries and developing countries are coming into an aging society. Growing elderly populations need more social welfare agencies and staffs which caused higher social costs. More and more consumer robots are designed for elderly service to replace health care staffs. Large number of current researches show the potential of robot served in elderly care service, J. Gallego-Perez summarized the methodological approaches to evaluate robot in elderly care, and listed the challenges to overcome in the present and the future [1]. [2] shows some challenges of elderly care robot in domesticated environments such as house work and human assistance activities. A fetch task was proposed using 3D spatial languages in human robot communication [3] [4]. Similar to our "HomeMate" robot, it supposed to work in the environment of domestic, hospital or elderly care center with the services of taking order, errand, medicine delivery and video chatting, etc. Our research mainly focuses on a natural calling gesture based taking order service that can locate user and deliver user required objects in a cluttered and noisy environment.

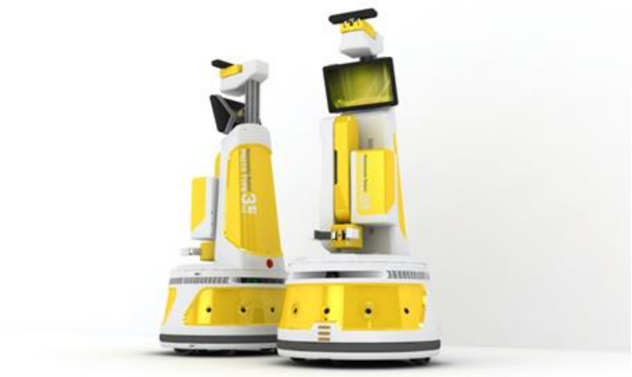

Figure 1. "HomeMate" a service Robot developed for Elderly Care

Recently, sensors such as Microsoft Kinect and ASUS Xtion pro were developed for somatic games; these sensors can provide gesture recognition, voice recognition and a new way to control games. Meanwhile, researchers also applied these kind of sensors for Human Robot Interaction (HRI) case [5] [6] [7]. The gesture recognition based HRI approaches were used for robot control [8] [9]. When robot is to provide a service, a moving senor based gesture recognition algorithm is proposed in [10]. Our goals are to design a natural calling gesture based taking order service scenario for elderly care robot.

The rest of the paper is organized as following: in section II, overview of our calling gesture based taking order scenario for elderly service robot is introduced. Section III described some disadvantages of the calling gesture recognition using Kinect for taking order services. The two modes of our calling gesture recognition approaches are introduced in section IV. Experiments and evaluations will show in section V, final section is conclusions.

II. Overview of Calling Gesture based Taking Order Service for Elderly Care Robot

A. System Structure

Our system (Fig. 2) includes "HomeMate" elderly service robot, a smartphone for remote interaction, and set of wireless routers for indoor localization. "HomeMate" is designed as a cognitive consumer robot developed for elderly care with services such as: errand, taking order, game playing, video chatting, as well as medicine alarm. It is equipped with Microsoft Kinect RGBD camera; Bumble bee 2 Stereo camera;

Xinshuang Zhao is with the Intelligent Systems Research Institute, School of Information and Communication Engineering, Sungkyunkwan University, south Korea (e-mail: ztchit@hotmail.com).

Ahmed M. Naguib is with the Intelligent Systems Research Institute, School of Information and Communication Engineering, Sungkyunkwan University, south Korea (e-mail: ahmed.m.naguib@gmail.com).

Sukhan Lee†, the Corresponding Author, is with the Intelligent Systems Research Institute, School of Information and Communication Engineering and the Interaction Science Department of Sungkyunkwan University, South Korea (e-mail: lsh@ece.skku.ac.kr)

Vision Pan/Tilt Module; Cognitive Vision PC; StarGazer Indoor Localization System, 5DOF manipulator and grasping end effector; laser sensor and Linux based robot PC.

B. *Overview of Taking Order Scenario*

Taking order scenario is the final state of errand service scenarios. After elderly call for a service, “Homemate” performs the required task by searching for the target object, localizing and obtaining it, performing additional service-related tasks (such as: filling juice from nearby juice dispenser), and finally, localizing elderly and handing the object to him/her. Our taking order scenarios (Figure 2) include five steps: 1. Find the correct room in which elderly is located; 2. Elderly searching and localization; 3. Approaching trajectory and angle correction; 4. Object handing; 5. Robot moves back to the charging place.

With development of smart phone technologies, applications for Smartphone now exist in many fields in our life. Wireless network technology makes a faster network connection between cell phone and other components. Smartphone nowadays becomes another useful tool for human robot interaction. In our scenario, a set of wireless routers are randomly distributing within the workspace. Elderly, randomly located in any of workspace rooms, call the robot using smart phone for service. After robot prepare the required order, Smartphone computes room number using a simple classifier of signal strengths of distributed routers. Assuming elderly is carrying his smart phone, it then sends this room number to the robot. With a prior knowledge of each room’s architecture, robot goes to predefined searching spots and start scanning areas of the room that elderly is most likely to be in (Fig. 3). Optimized gesture recognition, which is being proposed in this paper, is being utilized to locate elderly among cluttered and crowded environment. Elderly can be standing or sitting on a chair or sofa. Elderly forward body vector is computed and a navigation trajectory starts to approach him. Finally, a user verification and angle correction is performed and object handing process is carried out to deliver the target object. After that, scenario is finished and robot heads back to initial charging zone.

Proposed calling gesture recognition, used in the above taking order scenario, includes two modes: 1) Kinect skeleton based gesture recognition approach: in this procedure, user segments and skeleton information is generated by Microsoft Kinect, a verification process of Haar like feature based face detection is applied on each skeleton, and several calling gestures are defined for natural using; 2) Octree segmentation based calling gesture recognition [11] which is used as supplementary. In the following section, we will state the problems of using Microsoft Kinect skeleton based gesture recognition, and explain how our proposed algorithms provides an active solution to these problems.

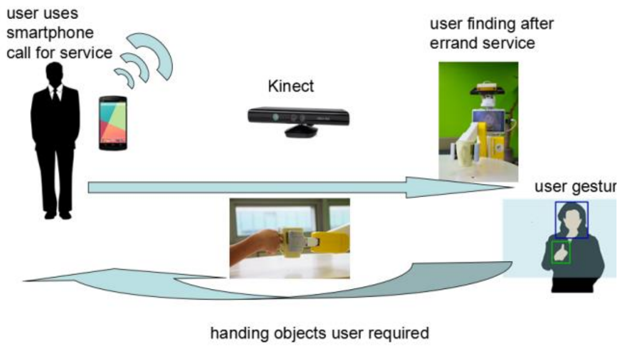

Figure 2. System structure

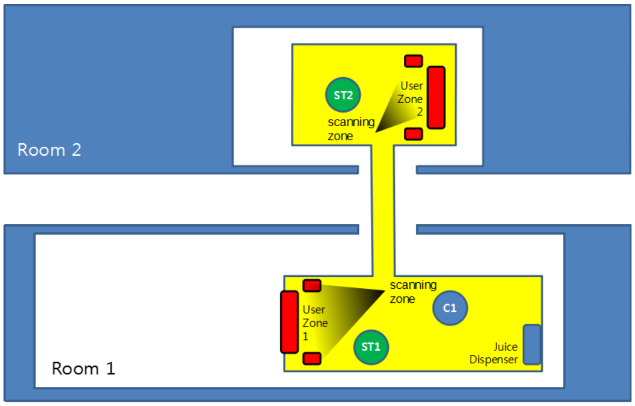

Figure 3. Scenario of taking order service for elderly care robot

III. Disadvantages of Original Kinect Skeleton Based Gesture Recognition

The field of view of Kinect has horizontal range of 57 degrees and vertical range of 43 degrees. Practically, the suggestion distance between user and sensor is 1.2 to 3.5 meters. By using infrared (IR) camera, depth image can be obtained, in the field of view of Kinect, maximum 6 users can be detected and two of them can be tracked with skeleton joints details. Kinect also can track users both in standing and sitting mode in real time. With these advantages of Kinect sensor, skeleton based gesture recognition attracted attentions of developers all over the world. At the beginning, Kinect sensor was designed for somatic games of Microsoft Xbox, nowadays, a lot of researchers used it for researches. By its real time and robust tracking capability, Kinect was also used in Human Robot Interaction field to control the robot. In our case, according to the scenario mentioned before, we applied Kinect sensor for natural calling gesture recognition in elderly service robot.

Due to the fact that Kinect was intended for an indoor gaming environment, it is designed to have a static pose in space looking at static background environment with foreground dynamic users. In a mobile service robot, however, both sensor and user are expected to have dynamic geometric poses that are constantly changing. This leads to number of problems in Kinect skeleton generation and tracking algorithms. These failure cases can be categorized as two main situations: A: skeleton generation failures; B: skeleton tracking failures.

A. *Skeleton generation failures*

Three failure cases occurred in this category.

a. False positives: Humanlike objects

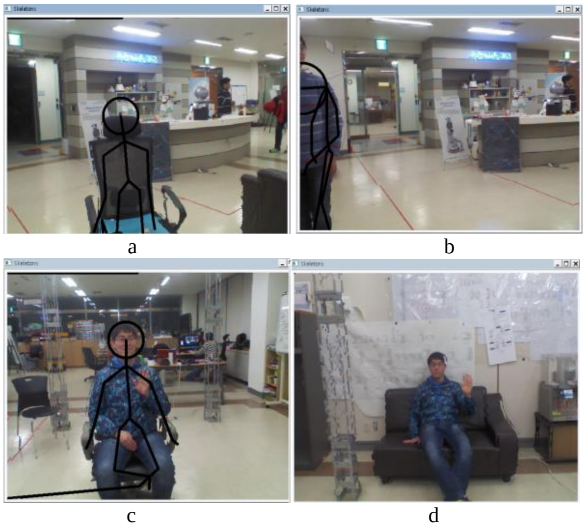

Figure 4. Skeleton errors by recognition

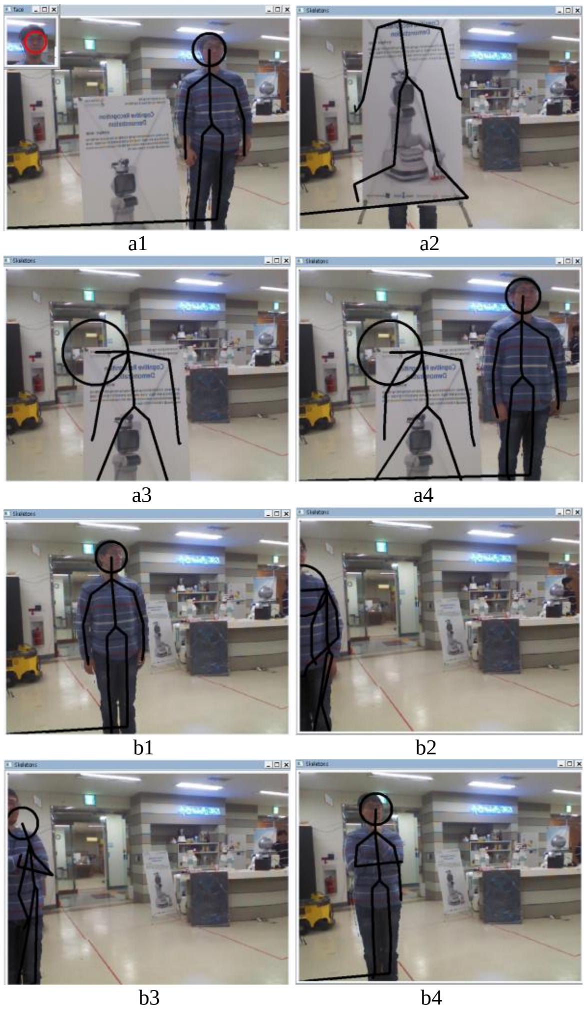

Figure 5. Skeleton errors by tracking

By Microsoft Kinect skeleton generation algorithms, users segments were created from depth image through infrared (IR) camera, shape configuration also used to identify user segments head and shoulder sizes, if the segments have a similar shape to a typical human, skeleton is generated. Unfortunately, this algorithm falsely generates skeleton for objects that have a generally similar shape outlines of a human form. Figure 4.a shows a normal arm chair that was detected as a human segment and skeleton of this segment was generated.

b. Skeleton distortion

If the user is standing or sitting at the edge of the Kinect field of view, generated skeleton will have noisy distortion (Fig. 4.b). This also occurs when the user is in contact with an object that has human-shaped limbs (Fig. 4.c). These noisy skeletons may trigger gesture recognition algorithm if not properly handled.

c. Missing a user

If the user is in contact with a geometric object that caused his head/shoulder shapes to take unusual forms, this segment will be rejected and skeleton won't be generated (Fig. 4.d). This happened in our experiment when a user was sitting on a typical leather sofa.

In normal case, users of our elderly service robot are elderly people. The skeleton errors happened in chair and sofa case cannot be ignored as elderly people are expected to be sitting on a chair or sofa.

B. Skeleton tracking failures

a. Fully drifted Skeleton: occlusion artifacts

The running environment of our elderly care robot is located at elderly center with crowd of people and clutter background. Users occluded by other people are highly expected. We simulated this sequence of tracking problem using a post board. User firstly appeared in the field of view of Kinect with the post board; user skeleton was initialized and started tracking. Then, the post board covered the user, the skeleton of user was then associated to the post board due to geometric constraint of tracking filter. User moved outside Kinect field of view and his skeleton was still assigned to the board. Finally user came back into the Kinect field of view, a new skeleton was generated for the user but wrong skeleton remained (Fig. 5 .a1 to .a4). Since human verification is only done at skeleton generation and not after that during tracking cycles, such tracking drifts are expected and an external verification process is required.

b. Partially drifted Skeleton: noisy joints and limbs

In taking order scenario, when robot is scanning user zone with synchronized motion, it is very likely that user skeleton might move to the edge of Kinect field of view and then moved back to center of the scene; Fig. 5 .b1 to .b4 described the error skeleton tracking sequences. User skeleton initialized, and user moved to the edge of the scene, skeleton became distorted, after user moved back to the center of the scene, wrong skeleton joints tracking persisted. For the case of application in elderly service robot, this tracking failure makes wrong detection or user missing, which happened for several times in our HomeMate robot full demo

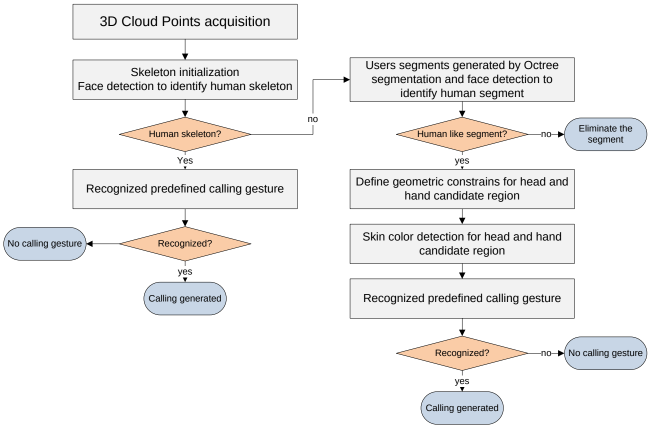
3D Cloud Points acquisition
Skeleton initialization
Face detection to identify human skeleton
Human skeleton?
no
Yes
Recognized predefined calling gesture
No calling gesture
Recognized?
yes
Calling generated
Users segments generated by Octree segmentation and face detection to identify human segment
Human like segment?
no
Eliminate the segment
yes
Define geometric constrains for head and hand candidate region
Skin color detection for head and hand candidate region
Recognized predefined calling gesture
Recognized?
no
No calling gesture
yes
Calling generated

Figure 6. Overall process flow chart of two mode calling gesture recognition

IV. Two Modes Gesture Recognition and Handing Approach

According to the two categories of skeleton failures discussed above, we are introducing an external verification process to only accept right skeletons and discard others. False positive generation of skeletons by human like objects can be discarded after applying a simple face detection verification process. Skeletons initialized without human face are considered as invalid skeletons. Skeleton distortion at the edge of field of view can be discarded too by certain angle threshold related to camera plane. In case of a false joints generation or missing skeleton, our proposed framework will switch to Octree based calling gesture recognition (Fig. 6). In Octree-based mode, human is recognized every frame and gesture is recorded. Overall performance is slightly slower, but artifacts generated from drifted tracking are avoided.

A. *Skeleton-based Calling Gesture Recognition with Face Verification*

When robot was scanning user zone, 3D point cloud acquisition of Microsoft Kinect sensor is performed. Simultaneously, Kinect SDK continuously carry out 3 processes: 1) recognize and locate new human segments, 2) track segments that were recognized before, and 3) generate skeletons for these segments. Since this SDK is not available for open source, we had to apply external components to avoid failure cases. In order to distinguish the skeletons whether were human or other human like objects, Haar feature based Adaboost face detection algorithm is applied on each skeleton segment, skeleton segments with human faces are retained while skeletons without human faces are discarded. There are 20 joints associated with each skeleton. In order to speed up face detection algorithm, only ROIs of circles centered at each head joint with radius of head-shoulder-center distance are used for face detection. Finally, Pre-defined natural calling gestures are identified using spatial-temporal motion of hand, wrist, elbow and shoulder joints of both left and right arms. If calling gesture is recognized, user's location information is used to approach user and robot starts object handing to users. The categories of skeleton based calling gestures' type will be discussed later.

B. *Octree-based Calling Gesture Recognition*

The Octree based calling gesture recognition is activated when skeleton generation fails. As mentioned above, 3D point cloud acquisition is continuously performed. Processing such dense cloud, however, is very computationally demanding specially in targeted crowded and highly cluttering environment. Instead, we first re-represent the scene using a sparse spatial representation of voxels, called "Octree cells". This representation dramatically reduce the sheer number of geometric building blocks we have to deal with, which leads to a rapid system that can coup with calling gesture speed we are expected to identify. Using this representation, a fast objects segmentation is applied and a structured addressing system and geometrical shape descriptor for each geometric segment is generated.

These segments are generated not only for humans, but also for other objects in workspace and background such as tables, sofa, chairs … etc. Haar feature based face detection algorithm is applied in each octree segment to identify which segment is a human segment and discard others. Since calling gesture is typically operated by user's hand, the head and hand candidate region are identified by cube regions. For each segment, a 30cm x 30cm x 30cm region around center of detected face is defined as head candidate region. A hand candidate region is then defined in a 50cm thick shell box region surrounding head candidate region. Skin color detection is applied in hand candidate region. HSV color space was used for skin color detection. Pixels are first separated to chromatic and achromatic groups in SV space. Chromatic pixels are then represented in HS space and quad-tree clustering algorithm clusters them into color segments. Each segment is then represented as a point in a 7D feature space by extracting: mean H, mean S, Eigenvector angel, Eigen value, Eigen value ratio, Weight ratio, and Scattering index. In order to match skin color of candidate region to skin color database using this 7D feature space, a multivariate density is computed from each segment point to a hyper ellipsoid database and likelihood probability is computed using Bayesian theory. We discarded region with probabilities below a certain threshold, and assign high skin color probability region in hand candidate regions as hand.

Figure 7. User study by inviting elderly people from elderly center

C. User Study Oriented Natural Calling Gesture Designs

Our calling gesture was designed for elderly users to call for service by service robot. Consider the elderly mobility and non expert capability, calling gestures are supposed to be as natural and simple as the system allows. In our study, a user study was held by capturing natural gestures carried out by random 30 elderly people from elderly centers (Fig. 7). These gestures are captured in the absence of robot, and in the presence of robot. Our study concluded 3 main types of gestures that covers majority of recorded gestures (Table I). These three types of calling gestures can be recognized by relative distance of human body in both modes.

D. Handing Object to Users

After users were found by calling gesture recognition, the handing objects of service, such as: cup of juice, water, milk box, and medicine grasped by robot in errand service should be delivered to users' hand correctly. When robot moved to users, by navigation errors, the handing angle toward user might be incorrect, face detection applied again on user segment to adjust approaching angle toward user. Robot then initiate manipulator pose for object handing and wait for user to take the object. There are many ways to identify whether user has taken the target object or not, such as: pressure sensor, tactile sensor, force feedback sensor … etc. we, however, relied on our robust octree representation to perform a visual-based object handing. This is simply by carrying out 2 steps. 1) Once an octree segment connects to robot's manipulator segment and then disconnects (Fig. 8), there is a high possibility object is taken. 2) After previous sequence, a verification process of existence of object in hand using depth information of sensor is performed. If there is a hollow in manipulator end effector, object has been taken successfully. Robot can then retracts, initialize his arm and move back to charging zone.

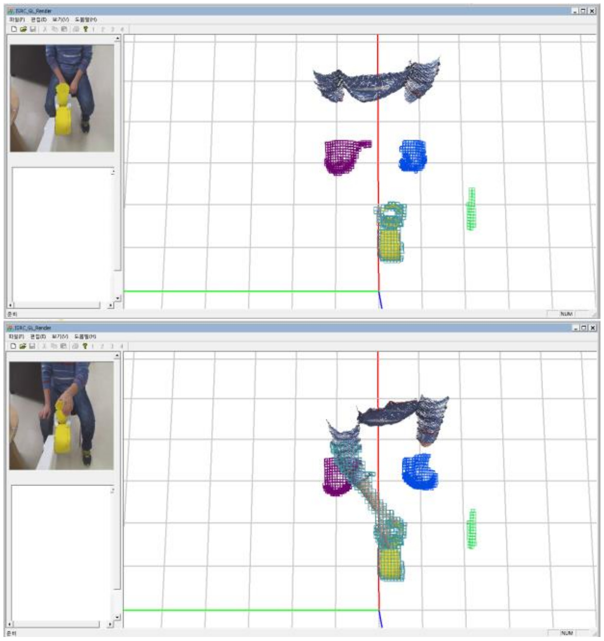

Figure 8. Handing object to user

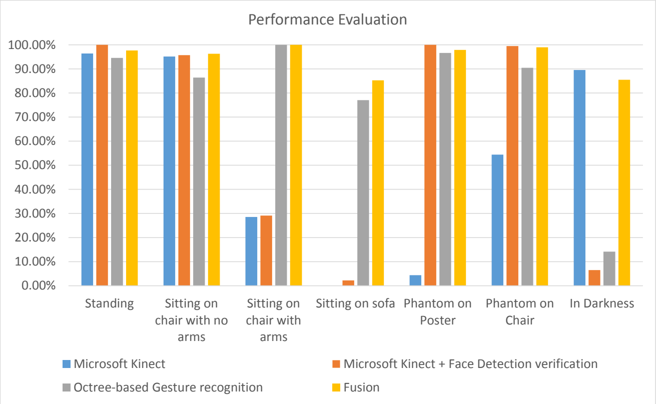

Figure 9. Performance evaluation

V. Experiments and Evaluation

In order to evaluate our calling gesture based taking order service, we mainly focused on the performance of our two modes calling gesture. Section III shows many disadvantages of Kinect skeleton based gesture recognition, our experiment here also compare the performance of original Kinect skeleton based gesture recognition method and our improved gesture recognition system.

For evaluating proposed method and performance, our "HomeMate" robot with its onboard Kinect Sensor were used as our experimental platform. The robot main computer is a tablet PC with Intel(R) Core (TM) i7-264M CPU @ 2.8G Hz, RAM of 4.00 GB. Our software was developed in C++.

In our evaluation procedure, we ran four different types of calling gesture recognition groups combined the two modes at the same time to check performance as: Original Microsoft skeleton based gesture recognition; Microsoft skeleton with face detection (mode 1); Octree based gesture recognition (mode 2) and fusion of mode1 and mode2. Seven different testing situations included standing, sitting on chair with no arm, sitting on arm chair, sitting on sofa, phantom on poster, phantom on chair, and in darkness were tested with the four types of gesture recognition methods respectively. We ran the system and test calling gesture 20 times in each situation with different position of sensor and user. The performance is shown in table II and Fig. 9 and evaluation images are in Fig. 10, demonstration video is shown in [12].

VI. Conclusion

Experimental results clearly show the advantages and disadvantages of each approach. While MS Kinect provides a fast skeleton that is easy to process, it fails to verify it constantly and tracking may drift away. Octree approach may provide a robust way to extract user regions; it fails, however, to operate under severe illumination conditions due to failures in skin color detection.

Our results conclude that while each approach may be good, a fusion of them as multiple evidences provides the best results and overcomes the weaknesses of their individuals.

Our future work includes using an open source skeleton extractor so that we can fuse these methods on a much lower

TABLE I. CALLING GESTURES

Calling Gesture Types	*Gesture Motion*
Lift arm	Lift arm in front of body and hold
Shake arm left and right	Lift arm in front and shake arm left and right
Shake hand forward and backward	Lift hand and arm in front and shake hand forward and backward

level instead of dealing with Kinect SDK as a black box with skeletons that can only either be accepted or rejected.

ACKNOWLEDGMENT

This work was supported by Priority Research Centers Program through the National Research of Foundation of Korea (NRF), funded by Ministry of Education (NRF-2010-0020210). This work is also partially supported by MEGA science research and development projects, funded by Ministry of Science, ICT and Future Planning (2013M1A3A3A02042335).

REFERENCES

[1] Gallego-Perez, J., M. Lohse, and V. Evers. "Robots to motivate elderly people: present and future challenges." In *RO-MAN, 2013 IEEE*, pp. 685-690. IEEE, 2013.

[2] Beer, Jenay M., C. Smarr, Tiffany L. Chen, Akanksha Prakash, Tracy L. Mitzner, Charles C. Kemp, and Wendy A. Rogers. "The domesticated robot: design guidelines for assisting older adults to age in place." In *Human-Robot Interaction (HRI), 2012 7th ACM/IEEE International Conference on*, pp. 335-342. IEEE, 2012.

[3] Skubic, Marjorie, Zhiyu Huo, Tatiana Alexenko, Laura Carlson, and Jared Miller. "Testing an assistive fetch robot with spatial language from older and younger adults." In *RO-MAN, 2013 IEEE,* pp. 697-702. IEEE, 2013.

[4] Scopelliti, Massimiliano, Maria Vittoria Giuliani, and Ferdinando Fornara. "Robots in a domestic setting: a psychological approach." *Universal Access in the Information Society* vol.4, pp. 146-155, 2005.

[5] Murthy, G. R. S., and R. S. Jadon. "A review of vision based hand gestures recognition." *International Journal of Information Technology and Knowledge Management* 2.2, pp. 405-410, 2009.

[6] Srikanth, M. V. V. N. S., et al. “Survey on Various Gesture Recognition Techniques for Interfacing Machines Based on Ambient Intelligence”. No. arXiv: 1012.0084. 2010.

[7] Weinland, Daniel, Remi Ronfard, and Edmond Boyer. "A survey of vision-based methods for action representation, segmentation and recognition."*Computer Vision and Image Understanding* 115.2 pp. 224-241, 2011.

[8] Qian, Kun, Jie Niu, and Hong Yang. "Developing a Gesture Based Remote Human-Robot Interaction System Using Kinect." *International Journal of Smart Home* 7.4, 2013.

[9] Cheng, Liying, et al. "Design and implementation of human-robot interactive demonstration system based on Kinect." *Control and Decision Conference (CCDC), 2012 24th Chinese*. IEEE, 2012.

[10] Hanguen Kim; Soon-Hyuk Hong; Hyun Myung, "Gesture recognition algorithm for moving Kinect sensor," *RO-MAN, 2013 IEEE*, vol., no., pp.320, 321, 26-29 Aug. 2013.

[11] Zhao, Xinshuang, Ahmed M. Naguib, and Sukhan Lee. "Octree segmentation based calling gesture recognition for elderly care robot." In *Proceedings of the 8th International Conference on Ubiquitous Information Management and Communication*, p. 54. ACM, 2014.

[12] http://www.youtube.com/watch?v=3a9gAX4KSI8&feature=youtu.be.

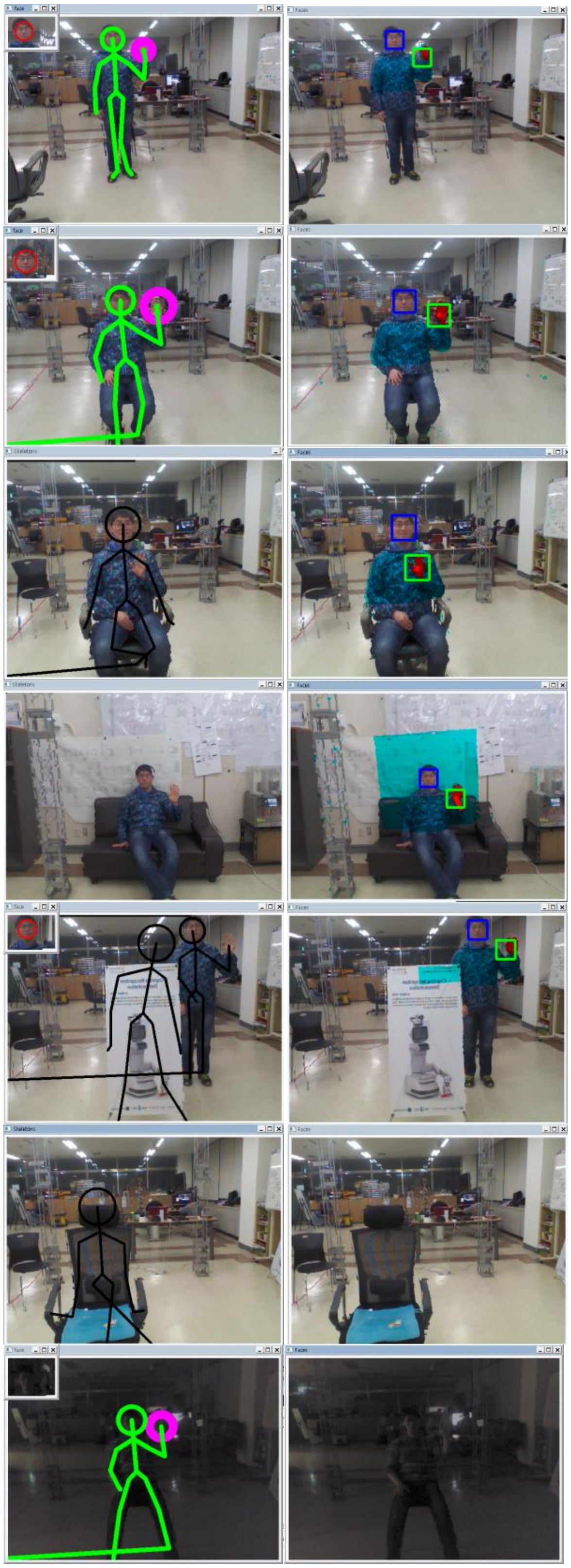

Figure 10. From top: Evaluation of standing, sitting on chair with arm, sitting on arm chair, sitting on sofa, phantom on poster, phantom on chair, and in darkness