2013 10th International Conference on Ubiquitous Robots and Ambient Intelligence (URAI)
October 31-November 2, 2013 / Ramada Plaza Jeju Hotel, Jeju, Korea

Ubiquitous Human Perception for Real-Time Gender Estimation

Anara Sandygulova, David Swords, Sameh Abdalla, Mauro Dragone and Gregory M.P. O'Hare

CLARITY: Centre for Sensor Web Technology, University College Dublin, Dublin 4, Ireland
(E-mails: {anara.sandygulova, david.swords@ucdconnect.ie}, {sameh, mauro.dragone, gregory.ohare}@ucd.ie)

Abstract - In environments where robotic systems are deployed people often have different requirements for the robotic services and human-robot interaction methods. This paper presents a robotic system that exploits the advantages of ubiquitous perception in order to gather knowledge from multiple sensors and various modalities. This ubiquitous human perception will facilitate user profiling in order to support personalised services and individual human-robot interaction. This system combines ubiquitous smart sensing, methods of multi-modal human perception and existing human recognition algorithms from the field of biometrics to collectively work towards a real-time, robust and scalable solution for *gender* estimation.

Keywords - Human-Robot Interaction, Ubiquitous Perception, Robotics, Smart Environments.

1. Introduction

As robotics research is steadily becoming ubiquitous [?] by moving into the smart environments, system's knowledge and cognition are no longer confined to the individual robot. It is rather distributed in the environment exchanging knowledge with the objects, inhabitants and tools via spatially placed sensors, mobile robots and technologies from the fields of ubiquitous computing and ambient intelligence. Such robotic solutions promise to provide affordable and scalable real-world applications to support our elderly to live independently, assist employees in smart hospitals or education institutions and, mainly cooperate in chores, as they become our new housemates.

A number of projects work towards human-robot interaction (HRI) within ubiquitous robotics to support elderly people to live independently [?, ?, ?]. These projects stress the importance of having a social interface between the smart environment and the user. However, the majority of ubiquitous robotics research has focused on the substantial technical challenges enabling their robots to seamlessly interact and operate in ubiquitous and smart environments, but avoided consideration of the wider social circumstances in which those services will ultimately reside. According to Davidoff *et. al.* [?] one of the main social characteristics that must influence the development of intelligent home services is that *"families are plural whereas most systems are singular"*. In addition, systems of ubiquitous robotics provide standard single user support capabilities and interaction techniques mainly because it creates fewer challenges [?]. However, smart environment is a social space shared by family, colleagues or temporal visitors, and systems need to consider multiple users and all the complexities this creates.

Cognitive robot companions need to demonstrate socialisation and personalisation in order to meet the social, emotional and cognitive needs of people they are sharing the common space with [?]. Individualisation is necessary due to the human nature: people have individual needs, preferences and personalities that a personalised robot would have to adapt to: one and the same robot will not fit all people. For instance, human-robot personalisation with Snack Delivery Robot [?] improved rapport, cooperation and engagement.

Contrary to previous research efforts, this work seeks to use the unique characteristic of ubiquitous robotic systems of multi-modal, multi-sensory perception to support multi-users HRI. Our system creates a user's profile by autonomously learning and understanding various individual characteristics to offer appropriate interaction and services. As a result of building profile for an individual user, the proposed system will support personalised interactions for multi-inhabitant smart homes, as well as other smart environments that are populated by many occupants including intelligent hospitals, schools, elderly centres, offices, and shopping malls.

The remainder of this paper is organised as follows: Section 2 introduces related work in the areas of HRI within ubiquitous robotics and multi-party HRI. Section 3 describes systems rationale while Section 4 examines the actual design and technical details of systems components. Section 5 details individual gender estimators used by three modalities and their fusion. Experiment and its results are discussed in Section 6, and Section 7 concludes our work.

2. Related Work

To date, ubiquitous robotics research has focused on making their robot social by integrating various HRI findings after careful consideration of a particular user group. For example, KSERA *("Knowledgable SErvice Robots for Aging")* investigates the integration of assistive home technology and service robotics to support older users in a domestic environment [?]. HRI of the KSERA system is established through small humanoid robot NAO employing non-verbal social cues such as joint attention via gaze and visual feedback. CompanionAble *("Integrated Cognitive Assistive and Domotic Companion Robotic Systems for Ability and Security")* project conducted user studies to investigate elderly needs and preferences: the exis-

978-1-4799-1197-4/13/$31.00 ©2013 IEEE

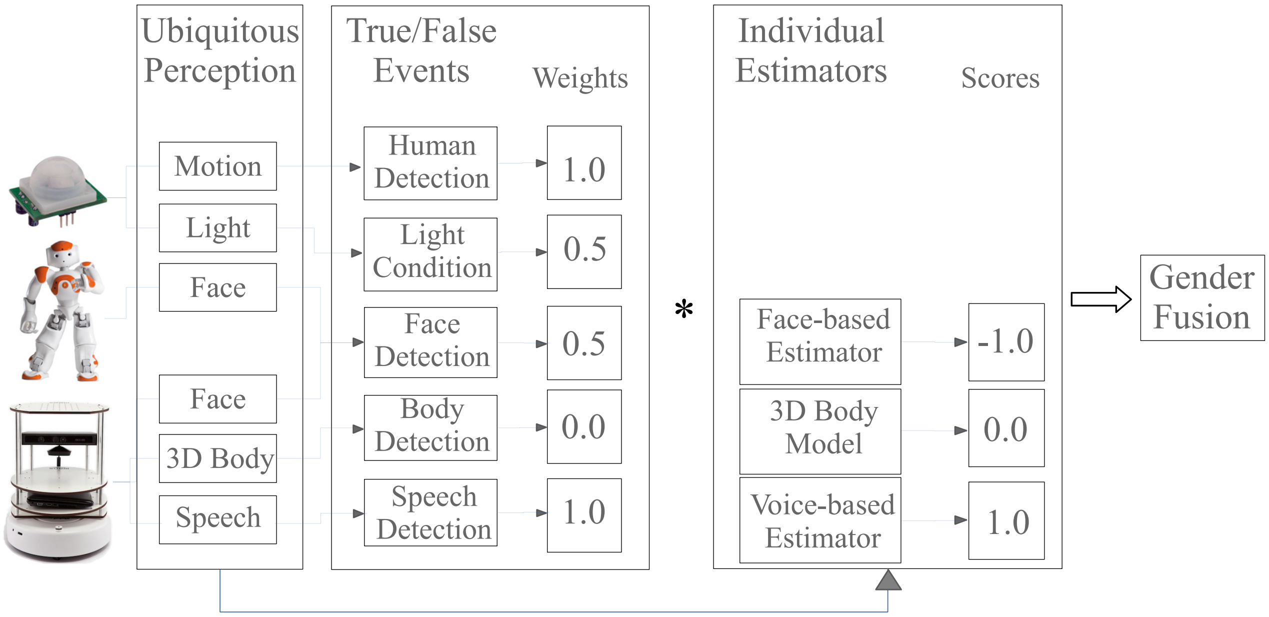

Fig. 1 Proposed Scheme.

tence and the design of a robot head plays a crucial role in the context of enabling and stimulating social interaction [?]. Due to these findings, a Hector robot has facial expressions to express emotions. The ACCOMPANY *("Acceptable roboticCs COMPanions for AgeiNg Years")* system consists of a robotic companion as part of an intelligent environment, providing services to elderly users in a motivating and socially acceptable manner to facilitate physical, cognitive and social assistance in everyday home tasks [?]. However, these and other projects that address independent living scenarios are *single user* systems, which offer *standard interaction* for all user groups and need to be *pre-programmed* for individual user needs.

Majority of real-world HRI is many-to-one communication. However, only a few research groups have addressed multi-party HRI for their systems. One example is a gesture-centric android system [?] that is able to adjust gestures and facial expressions based on a speaker's location or situation for multi-party communication. The speaker location is identified by face recognition and microphone position. The experiment was conducted with 1662 subjects interacting with Actroid-SIT android in a shopping mall in Japan. Another field experiment was conducted during a 6-day Fleet Week in New York with 202 subjects [?]. Groups of three people firstly trained the Octavia robot to memorise their soft biometrics information (complexity, height and clothes), then as the people tried to trick Octavia by changing their location, she could 90% of the time successfully identify people. Both systems successfully address multi-party HRI relying on multi-modal recognition of people: sound and vision. However, in case of changing environments and lighting conditions, the systems will likely show different results.

Many Ambient Assisted Living (AAL) applications gather context by using multiple modalities [?, ?] similar to our work, however they are usually focused on recognising the context of the user in terms of what activities he/she is engaged with (i.e. sleeping, in need of emergency, etc.), whereas our objective is to acquire a profile of the user to be used for HRI.

3. Ubiquitous Perception

A social intelligent environment should be able to understand and provide appropriate services and personalised interaction towards any user group. The challenge is that intelligent environment does not know its potential users and therefore it should be able to dynamically learn and classify new people's features and characteristics in order to adapt its interaction style and services. One of the main advantages of ubiquitous robotics is that of multi-sensory, multi-modal perception. In a traditional approach, standalone robot sensors mimic the functionality of human perception of vision and hearing: however, a ubiquitous robot can see and hear beyond its physical access and presence. For instance, physiological signals such as heart rate and blood pressure provide information about human's emotional state that may not otherwise be observable. What is more, a physical robot does not need to be collocated with a human to perceive, understand and react to the human activity in real-time. Perceptual capabilities of such applications need to be capable to adapt to dynamically changing conditions of lightning, colours, and objects in real environments since the range of sensor inputs is more extensive than most traditional robotic systems. Effective integration of multi-sensor, multi-modal data for human detection, recognition and tracking, body pose and movement, sound source localisation will be achieved though combining leading edge biometrics, computer vision, human-computer and human-robot interaction practices and findings within the ubiquitous nature of service robotics.

4. System Design

This section details the ubiquitous perception system starting from the design of its perception, its middleware and software and hardware components.

4.1 Perception Design

Figure **??** illustrates the proposed scheme of *gender* estimation with applied fusion of different human perception modalities. Ubiquitous perception here consists of readings from wireless sensors of light and motion nodes deployed in a room furniture, and of human perception information such as face, body and speech from mobile and humanoid robots. This information is used to trigger true/false events such as *human detection, light condition, body detection, face detection* and *speech detection* that shape the trustworthiness of a particular modality guaranteeing system's robustness via the fusion of individual *gender* estimators. Therefore, from no information about the user during continuous interaction with the robotic system, the system is able to perceive and learn user's characteristics over time.

A. Wireless Sensor Nodes

In order to support system's adaptability, for instance to adapt to changes in the environment such as lighting conditions and user presence, and increase overall system's robustness, we employ a wireless sensor network (WSN) comprising of nodes equipped with light and motion sensors. The readings from these sensors are communicated (over IEEE 802.15.4 ZigBee) by each WSN nodes to a base station node interfaced with a mini-pc, by using a dedicated software [**?**] that publishes those readings to the PEIS tuplespace. The readings from the light sensors are used to compute the trustworthiness of the vision-based sensing. This means that in a dark environment, the system does not rely on the vision modalities for estimating *gender* of the user. The readings from the motion sensors are used to deduce human presence when the vision sensing is not taken into consideration by the ubiquitous perception system.

B. Face Detection

Two robots are responsible for *face detection* in order to estimate *gender* of the user. The reason for having multiple sources to capture the face is to facilitate an extended and mobile field of view. In case the user is not facing or in range of a particular robot in a smart environment, the face data can be still captured by another robot. In addition, a range of facial expressions can cause errors in the face-based algorithms to estimate age, gender, ethnicity, etc., and the face captured at different points of time provides more data for more confident results of the face-based algorithms.

NaoQi programming framework of the humanoid robot NAO provides a vision module, ALFaceDetection, which the NAO can use to detect human faces. This module also provides an estimate for the position of each face detected in the frame grabbed by the NAO's camera, as well as a list of angular coordinates for a set of important face features. The Microsoft Face Tracking SDK engine is also used to analyse input from a Kinect camera to detect and track human faces in real-time. Once the face is detected, a *face detection* event triggers a real-time face-based gender estimation algorithm.

C. Speech Detection

Real-time streaming of audio from the Microsoft Kinect microphone array is used for *speech detection* purposes. Its Speech API provides communication with the microphone array. Once the audio signal from the microphone is significant enough to be speech, a *speech detection* event triggers a voice-based gender estimation algorithm.

4.2 Middleware

Software components, furniture and operating devices involved in the system are illustrated in Fig. **??**. The hardware components of the ubiquitous robotic environment are:

- a humanoid robot NAO[1];
- a TurtleBot[2] equipped with Microsoft Kinect[3];
- a smart furniture equipped with sensor nodes;
- a computer with a monitor acting as a Smart TV.

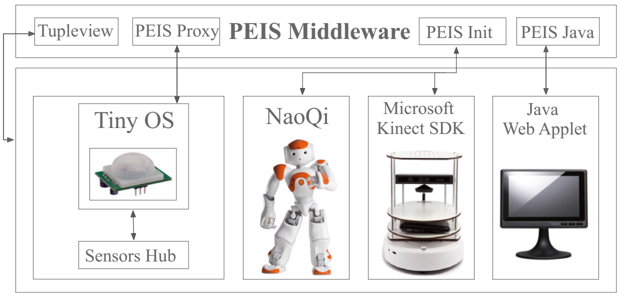

Fig. 2 System Components.

In order to effectively leverage highly heterogeneous resources of the ubiquitous robotic environment, zero configuration and interoperability among robots, sensors, actuators and traditional computers is ensured by using the PEIS kernel [**?**], a software suite previously developed as part of the Ecologies of Physically Embedded Intelligent Systems project [**?**]. PEIS includes a decentralised mechanism for collaboration between separate processes running on separate devices, which allows for automatic discovery, high-level collaboration through subscription-based connections. It also offers a shared tuplespace blackboard for the exchange and storage of tuples, (key-value pairs) used in associating any piece of data (and related meta-information, such as timestamps and MIME types), to a logical key. The PEIS kernel is written in pure

[1] www.aldebaran-robotics.com/eng
[2] www.willowgarage.com/turtlebot
[3] www.microsoft-kinect.com/kinect

C (with binding for Java and other languages) and with as few library and RAM/processing dependencies as possible to maximise compatibility with heterogeneous devices. Current ubiquitous robotic system is an extension of our previous work in [?].

5. Gender Estimation

This section describes three individual modalities for gender estimation (3D body model, face, and voice), and our method of data fusion.

5.1 Individual Estimators

Our work employs three open-source solutions from biometrics and computer vision research fields in order to demonstrate the concept of proposed ubiquitous perception system.

A. 3D Body Model Estimator

A motion capture device, such as Microsoft Kinect, is used for 3D body metrics that are particularly indicative of various demographics groups (i.e. gender or age). Gender or age group of a person in a field of view of the camera can be determined based on the metrics of a 3D body model. This includes the metrics such as the ratio of the arm length to the body height, the ratio of the shoulder width to the waist width and the hip width, the ratio of the body height to the head height or the ratio of the head width to the shoulder width, etc. Microsoft Kinect SDK provides the coordinates (X, Y and Z) of 20 skeleton joints. Calculating the distance between the coordinates of the point mesh and the ground floor of the depth map of the user does not provide reliable and flexible solution as the user can be sitting in a wheelchair, can be standing with bended knees or simply can lean to the side. Therefore, an adaptable algorithm to accurately determine person's *height* proposed by [?] is to calculate the *sum of the lengths*:

$$\begin{aligned} height = {} & length_{head-shoulderCenter} + \\ & + length_{shoulderCenter-spine} + \\ & + length_{spine-hipCenter} + \\ & + length_{hipCenter-kneeLeft/kneeRight} + \\ & + length_{kneeLeft/kneeRight-ankleLeft/ankleRight} + \\ & + length_{ankleLeft/ankleRight-footLeft/footRight} \end{aligned} \quad (1)$$

Since the Kinect estimates the head joint from the centre of the head, the total sum needs to add a few centimetres to the end of the head using the depth stream. Similarly to the *height* calculation, other useful body dimensions are extracted from the Kinect data to define user's body metrics. For the experiment described in this paper, our system uses 3D body model captured by the Kinect to estimate *gender*.

B. Face-based Estimator

In addition to the body measures, *face* is used as a source of information of someone's background from commonly examined demographics such as gender, age, and ethnicity to specific psychological meanings for each facial feature. OpenBR [?] is an open source biometric recognition framework for investigating new modalities, improving existing algorithms, interfacing with commercial systems, measuring recognition performance, and deploying automated biometric systems. OpenBR is designed to facilitate rapid algorithm prototyping, and features a mature core framework, flexible plug-in system, and support for open and closed source development. OpenBR also provides off-the-shelf algorithms for face recognition, age and gender estimation, which are integrated in our system. For the purposes of the experiment, the system uses the face-based *gender* estimation algorithm from OpenBR.

C. Voice-based Estimator

Another modality is *voice*. Gender is estimated with the help of the pitch detector algorithm [?]. The pitch, or fundamental frequency, of humans ranges from 75 Hz to 275 Hz. Adult males tend to occupy the range from 85 to 180 Hz, while adult females tend to occupy the range from 165 to 255 Hz [?]. Because of this distinct separation of pitch frequencies, gender recognition by speech analysis can be performed 97% accurately [?]. The algorithm performs its *gender* estimation in real-time within a few seconds.

5.2 Fusion

The fusion of multiple biometric traits could happen at various levels, such as the feature extraction level, the matching score level or the decision level. In practice, fusion at the matching score level is generally preferred due to the ease in accessing and combining the matching scores [?].

Each modality returns its gender score in the form of [-1, 1] where gender is either -1 or 1 for male and female respectively. For the person instance $x^{'}$, $p(g/x^{'})$ can be viewed as the score of gender g output from the individual gender estimators based on body, face or voice. Suppose $p_b(g/x^{'})$ is the body score, $p_f(g/x^{'})$ is the face score and $p_v(g/x^{'})$ is the voice score of gender g, then, the fusion score of g is calculated by

$$S(g) = w_b p_b(g/x^{'}) + w_f p_f(g/x^{'}) + w_v p_v(g/x^{'}), \quad (2)$$

where $0 \leq w \leq 1$ is the weight of the individual estimators. For the best-case scenario (during the artificial lighting conditions and when the data is present from all modalities), the weights $\{w_b, w_f, w_v\}$ are assigned to be {0.2, 0.5, 0.9} according to demonstrated precision during initial testing.

When data is missing for a given modality, the weight for that modality is assigned to be zero. The weights of the vision-based modalities varies according to lighting values and are set to be zero in a very poor lighting conditions close to complete darkness or in backlit situations. Similarly for the weight of the voice modality, w_v is set to be zero when the *speech detection* is false.

6. Experiment

6.1 Method

The experiment scenario is similar to the Smart TV functionality. It is fully autonomous and timed: the experiment starts as the human presence is detected in the environment by either a motion sensor equipped in the surrounding furniture, a full body detection or a face detection components of the system. Then, the NAO welcomes the user, introduces its name and asks for a participant's name. From the beginning, the system is collecting knowledge from different components in order to estimate the user's *gender* in order to launch a personalised video clip on a computer monitor. Time is logged in order to evaluate system's gender estimation performance over time as the availability of data from different modalities change. The experiment does not require any demographics information except for participant's gender and age.

6.2 Results

Sixteen people (age: M = 27.53, SD = 3.523; 8 males, 8 females) participated in the experiment. Ubiquitous robotic environment averaged 87.5% correct identification rate applying the ubiquitous human perception and the fusion of three modalities. This performance is very encouraging as throughout the experiment the system was able to collect necessary data for the individual estimators and perform its gender estimation and fusion in real-time. Due to the poor lighting conditions throughout the experiment and some background noise, two of the participants were mistakenly estimated. Both participants spoke very quietly and the background noise interfered with the voice-based estimator, which served as the most precise estimator in this experiment's conditions over such a short period of time. Table 1 outlines the gender estimation results obtained from each individual modality and after the applied methodology of the ubiquitous fusion.

Table 1 Gender Estimation Results.

Modality	# of Participants	Confidence (%)
3D Body Model	9	56.25
Face-based	11	68.75
Voice-based	14	87.5
Fusion	14	87.5

It might seem that the system did not improve from the voice-based gender estimation modality. However, in cases when the voice might not be available due to the background noise or if the user has not spoken yet, the system adapts to the available modalities for the knowledge it needs. The other diagram, Figure 2, for instance, shows the level of classification also when sound was not yet used (because the user hadn't spoken yet). Figure **??** illustrates an example of one participant's data to demonstrate a real-time gender estimation over time (30 seconds). As seen from the diagram, the system performed its gender estimation based on the 3D body model estimator, which performed the same estimation result till the end of the experiment. Face-based estimator produced various results over the whole duration of the experiment. The weight of this modality is also lower than the default one which is due to the lighting condition. Finally, as the voice-based estimator acquired speech, it influenced the resulting estimation score, which is normalised to 1.0.

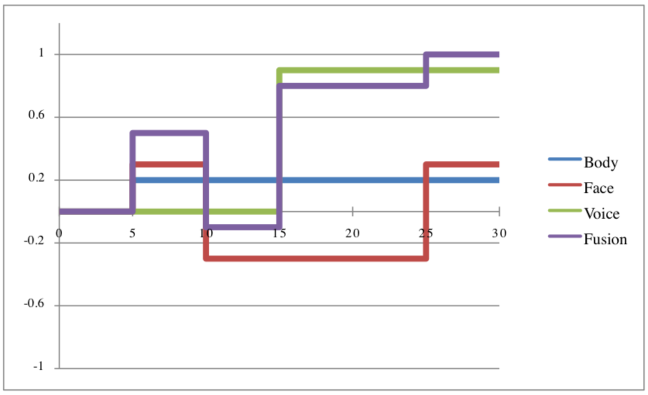

Fig. 3 Systems Fusion Performance Over Time.

7. Conclusion & Future Work

To date scant research has explored individuality in human-robot interaction and how this might empower the manner and form of interaction and assistive services. In this paper, we presented our approach to the development of a personalised robotic system that perceives human characteristics via robust and adaptable ubiquitous perception solution to estimate gender in real-time.

Our system's limitation is that it is not yet able to tackle multiple users at the same time but it handles multiple users that interact with the system one at the time. Expanding system's functionality to support multi-party HRI will be addressed in the future research. Our approach will support human body tracking, face tracking and sound source localisation while stationary mode of the robots used in the experiment provides ways to effectively localise themselves and users in the environment. Additionally, in the future work the system will exploit machine-learning techniques to investigate how to fuse multiple modalities without requiring ad-hoc and pre-programmed weights, but learning to assess a context-dependent trustworthiness of the different knowledge sources. To conclude, our work is at the stage in which the focus is to facilitate ubiquitous perception by effectively gathering information concerning the profile of the users aiming to facilitate personalised services and individual HRI.

Acknowledgement

This work has been supported by the Irish Research Council "Embark Initiative", in conjunction with funding from the EU FP7 RUBICON project (contract n. 269914), and Science Foundation Ireland (SFI) under grant 07/CE/I1147.

References

[1] Jong-Hwan Kim, “Ubiquitous Robot: Recent Progress and Development”, *SICE-ICASE International Joint Conference*, Busan, Korea, pp. I-25 - I-30, Oct. 2006.

[2] R. H. Cuijpers, “Ksera presentation,” Lecce, Italy, 2011.

[3] A. Badii, C. Huijnen, H. Heuvel, D. Thiemert, and H. Nap, “Companionable: An integrated cognitive-assistive smart home and companion robot for proac- tive lifestyle support,” *Gerontechnology*, vol. 11, no. 2, 2012.

[4] S. Bedaf, G. Gelderblom, F. Guichet, I. Iacono, D. Syrdal, K. Dautenhahn, H. Michel, P. Marti, F. Amirabdolahian, and L. Witte, “Functionality of service robotics for aging-in-place: What to build?” *Gerontechnology*, vol. 11, no. 2, 2012.

[5] Scott Davidoff, S., Lee, M. K., Dey, A. K., and Zimmerman, J. (2006). “Socially-aware requirements for a smart home.” *In Proceedings of the 2006 IEEE Conference on Intelligent Environments (IE 2006)*, 45-48.

[6] G. Amato, M. Broxvall, S. Chessa, M. Dragone, C. Gennaro, R. Lopez, L.P.Maguire,T, M.McGinnity, A.Micheli, A.Rentera, G.M.P. O’Hare, and F. Pecora, “Robotic UBIquitous COgnitive Network”. *In ISAmI 2012*, pp. 191-195.

[7] K. Dautenhahn, “Robots we like to live with?! - a developmental perspective on a personalized, lifelong robot companion,” *in Robot and Human Interactive Communication, 2004*. ROMAN 2004. 13th IEEE International Workshop on, 2004, pp. 17-22.

[8] M. K. Lee, J. Forlizzi, S. Kiesler, P. Rybski, J. Antanitis, and S. Savetsila, “Personalization in HRI: a longitudinal field experiment,” *Human-Robot Interaction (HRI), 2012 7th ACM/IEEE International Conference on*. New York, NY, USA: ACM, 2012, pp. 319-326.

[9] Y. Kondo, K. Takemura, J. Takamatsu, and T. Ogasawara, “A Gesture-Centric Android System for Multi-Party Human-Robot Interaction,” pp. 133-151, Jan. 2013. *Journal of Human-Robot Interaction.*

[10] Eric Martinson, Wallace Lawson, and Greg Trafton. 2013. “Identifying people with soft-biometrics at fleet week”. *In Proceedings of the 8th ACM/IEEE international conference on Human-robot interaction (HRI ’13).* IEEE Press, Piscataway, NJ, USA, 49-56.

[11] Anastasiou, D. (2012). “Gestures in Assisted Living Environments.” *Gesture and Sign Language in Human-Computer Interaction and Embodied Communication*. Springer Berlin Heidelberg, 2012. 1-12.

[12] Giuseppe Amato, Mathias Broxvall, Stefano Chessa, Mauro Dragone, Claudio Gennaro, and Claudio Vairo, “When Wireless Sensor Networks Meet Robots”, *7th International Conference On Systems and Networks Communications (ICSNC)*, Lisbon, Portugal, 18-23 November 2012, pp. 35-40.

[13] W. K. M. Broxvall, B.S. Seo, “The peis kernel: A middleware for ubiquitous robotics,” *In Proc. of the IROS-07 Workshop on Ubiquitous Robotic Space Design and Applications*, San Diego, California, 2007.

[14] A. Saffiotti and M. Broxvall. “PEIS Ecologies: Ambient intelligence meets autonomous robotics”. *In Proceedings of the International Conference on Smart Objects and Ambient Intelligence (sOc-EUSAI)*, pp. 275-280, Grenoble, France, 2005.

[15] Sandygulova, A.; Swords, D.; Abdel-Naby, S.; O’Hare, G.M.P.; Dragone, M., “A study of effective social cues within ubiquitous robotics,” *Human-Robot Interaction (HRI), 2013 8th ACM/IEEE International Conference on*, vol., no., pp.221-222, 3-6 March 2013.

[16] Vangos Pterneas. “Kinect for Windows: Find user height accurately”. Code Project. 9 May 2012. Web. 29 May 2013.

[17] Josh Klontz, “OpenBR Open Source Biometric Recognition and Beyond” (slides). Retrieved from http://openbiometrics.org/slides.pdf. February 17, 2013.

[18] Parker, Radford. “Real-Time Kinect Player Gender Recognition using Speech Analysis.”

[19] Geng, X., Fang, E., Smith-Miles, K., 2011, “Fusion of face and voice for automatic human age estimation”, *In Proceedings of the 3rd International Conference on Computer Design and Applications*, IEEE, Singapore, pp. 311-314.