Multi-modal Person Localization And Emergency Detection Using The Kinect

Georgios Galatas
Computer Sci. and Eng. Dept.
University of Texas at Arlington
Arlington, TX, USA
IIT, NCSR Demokritos
Athens, Greece

Shahina Ferdous
Computer Sci. and Eng. Dept.
University of Texas at Arlington
Arlington, TX, USA

Fillia Makedon
Computer Sci. and Eng. Dept.
University of Texas at Arlington
Arlington, TX, USA

***Abstract*—Person localization is of paramount importance in an ambient intelligence environment since it is the first step towards context-awareness. In this work, we present the development of a novel system for multi-modal person localization and emergency detection in an assistive ambient intelligence environment for the elderly. Our system is based on the depth sensor and microphone array of 2 Kinect devices. We use skeletal tracking conducted on the depth images and sound source localization conducted on the captured audio signal to estimate the location of a person. In conjunction with the location information, automatic speech recognition is used as a natural and intuitive means of communication in order to detect emergencies and accidents, such as falls. Our system attained high accuracy for both the localization and speech recognition tasks, verifying its effectiveness.**

Keywords-localization; multi-modal; Kinect; speech recognition; context-awareness; 3-D interaction

I. INTRODUCTION

An assistive ambient intelligence environment is a smart space that aids the inhabitants with its embedded technology. For achieving this goal, activity recognition and emergency detection, performed in a natural and unintrusive way, are of utmost importance. The most decisive step for effective activity recognition is accurate and robust person localization. By utilizing the location of the person in a domestic setting, the related activity can be derived. In addition, an elderly-oriented assistive environment must also be able to detect emergencies and accidents, such as falls, in order to issue a distress signal. This additional role increases requirements for information redundancy and reliability. Our novel system uses information from multiple sensors in order to ensure reliable localization of the inhabitant as well as emergency detection, offering speech recognition as a means of natural interaction.

Applications that rely on localization such as surveillance and monitoring of assistive daily living (ADL) commonly use video cameras as an affordable and abundant source of information. Many approaches based on either a single camera or multiple cameras have been proposed in the literature. In single camera setups, discriminative appearance affinity models [1] and level-set segmentation [2] have been used for tracking, while other approaches based on tracking-by-detection exist [3, 4]. In multi-camera setups, stereo-vision is employed in order to introduce depth perception. In [5], color histograms of the detected persons are used, while in [6] heuristic and probabilistic tracking is used to determine the location of a person in the 3-D space. Nevertheless, the segmentation and tracking problems can be very challenging, thus hindering the system's reliability in a camera-only setup. In an attempt to improve performance, multi-modal person localization has become a significant research area in recent pervasive assistive applications. Common approaches combine cameras with microphone arrays and other sensors for localization and activity detection [7], using particle filtering for data fusion. In addition to activity recognition, some systems incorporate emergency detection and most notably fall detection. Fall detection has triggered the interest of researchers, since falls account for over 75% of domestic accidents for adults over 75 [8]. Furthermore, over 30% of adults over 65 fall at least once a year [9], making them the most common cause of injury death [10] with a direct cost of $30 billion [11]. Although the vast majority of fall detection systems use solely cameras [12, 13], some systems use a combination of sensors [14].

Nevertheless, the use of cameras in all the aforementioned implementations can be considered intrusive when used domestically. Furthermore, fall detection systems do not account for other hazards that may arise in a domestic setting and may require additional modalities to increase reliability. However, our system uses the depth sensor of a Kinect device as the main source of localization information. This active sensor is able to accurately measure the position of the person in the 3-D space. At the same time, the video information is not captured, making this approach less intrusive than using video cameras. In addition, the Kinect device incorporates a microphone array capable of localizing sounds. In our system, we use 2 Kinect devices to capture the audio signal for both improving localization by introducing an additional modality as well as enabling natural interaction by means of speech recognition. Speech recognition is also used in order to detect emergencies independently of the localization module.

In the following sections we will present the architecture and operation of our system for the person localization and emergency detection tasks, the experimental setup and finally our concluding remarks.

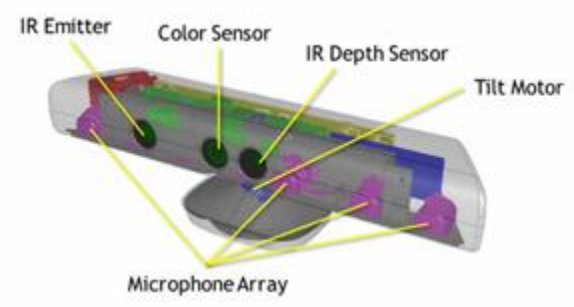

Figure 1. The Microsoft Kinect device.

II. The Kinect Device

The Microsoft Kinect (fig. 1) is a new device mainly used for gesture recognition. It is based on the PrimeSensor design [15] and it incorporates a color camera, a depth sensor and a microphone array. Depth images are acquired using the structured light technique. According to this method, a laser beam passes through a grating, and is split into different beams. The beams are then reflected from an object in the device's field of view (FOV) and captured by an infra-red sensor, making it possible to calculate the distance of the object using triangulation [16]. The microphone array is comprised of 4 microphones, enabling sound source localization. For our application, we implemented the least intrusive setup possible by capturing data only from the depth sensor and the microphone array, without capturing the actual color video data.

III. System Architecture

The architecture of our system is modular, comprising of 3 main components as shown in fig. 2. Communication between the modules is based around the Joint Architecture for Unmanned Systems (JAUS) [17], originally developed by the U.S. Department of Defense, to govern the way that unmanned systems are designed. The user datagram protocol (UDP) is used for inter-module communications, which increases the level of interoperability, allowing new software modules to be easily integrated in the system or existing modules to be installed on different systems. Input is provided by 2 Kinect devices. One of them is considered as primary, capturing both a stream of depth images and audio, while the secondary captures only audio for performing sound localization. Interfacing with the Kinect is carried out using the MS software development kit (SDK) v1.0 [18]. The 3 modules 1) skeletal tracking based localization, 2) audio localization and 3) automatic speech recognition (ASR) are described in detail in the following paragraphs.

A. Skeletal Tracking Based Localization Module

Skeletal tracking is used in our system in order to detect and track a person in the FOV of the sensor, as s/he moves in the smart space and it was implemented using the MS Kinect SDK. Initially, the moving person is detected, then her/his center of mass is determined and finally a skeletal model is fitted. The detected skeleton has a unique identifier for a specific session and is defined by the 3-D coordinates of its 20 joints $< X_{di}, Y_{di}, Z_{di} >$, expressed in meters.

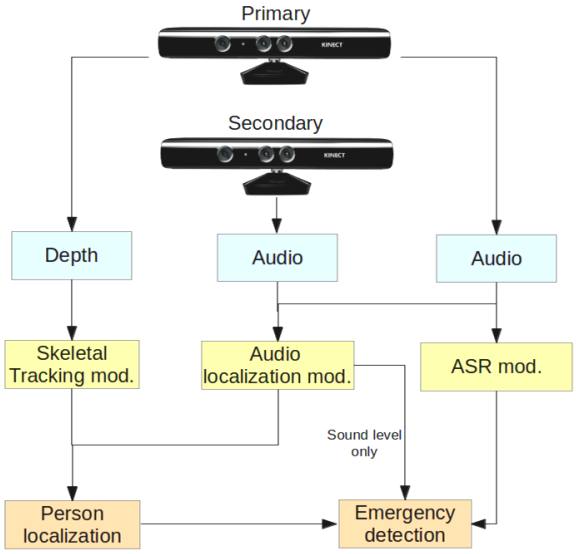

Figure 2. System architecture showing the 3 modules and 2 operation modes.

Each joint can be at any of the three associated states: 1) tracked, 2) not-tracked and 3) inferred. Furthermore, two kinds of filters are applied to the joint coordinates due to the nature of the captured data, 1) high frequency jitter and 2) temporary spikes rejection. Although the infrastructure for tracking the joints of 2 skeletons and the center of mass of 4 additional people exists, the main scope of our system is to monitor an elderly inhabitant of an assistive environment when not supervised, so at most 2 tracked skeletons are considered. Localization using such skeletal tracking is very accurate and unintrusive since we only utilize the coordinates calculated from the depth sensor feed. A visualization of the operation of the skeletal tracker is shown in fig. 3.

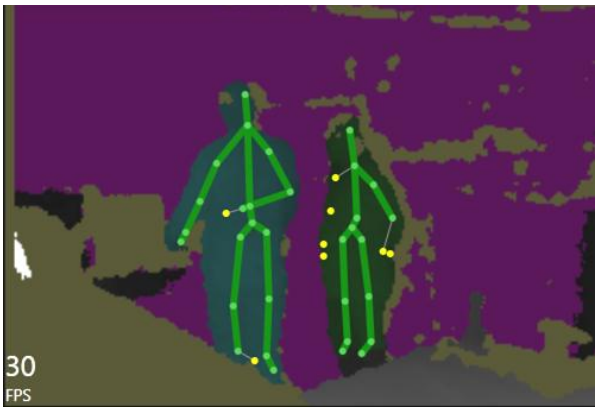

Figure 3. Skeletal tracking example.

B. Audio Localization Module

The microphone array of the Kinect is comprised of 4 supercardioid microphones that drive 24-bit ADC's. The frequency response of the microphones is tailored for human speech and their directivity is relatively stable for these frequencies (1-7 kHz). Sound source localization and beamforming are applied to the audio signal in order to determine the angle of the sound source in relation to the

device and acquire the audio signal from that particular direction (fig. 4). The returned values are the sound source angle in degrees in relation to the axis that is perpendicular to the device, and a confidence level of the reported angle.

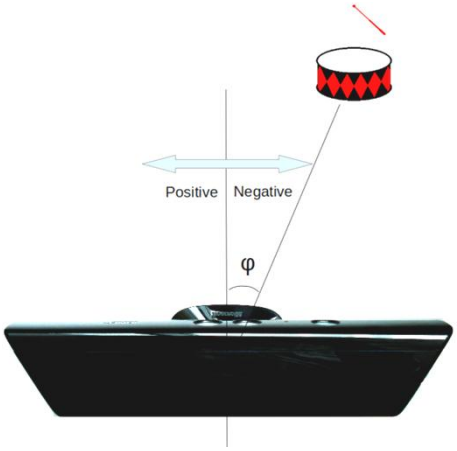

Figure 4. Kinect sound source localization.

Nevertheless, one Kinect is only capable of providing the angle of the sound source but not its distance, hampering localization accuracy. Therefore, we introduce a second Kinect to our system that is used solely for sound source localization (fig. 5). The second unit also provides an angle for the source of the sound, which can be used in combination with the previously obtained angle for accurate localization through triangulation. In order to do so, we need to obtain some data concerning the placement of the sensors. More specifically, let L be the distance between the two devices, A and B. Also, let θ_A, θ_B be the angle between the wall and the axis perpendicular to device A and B respectively. This angle should optimally be 45 degrees to maximize coverage assuming the devices are mounted at the corners of the same wall in a square room. Assuming there is a sound source S detected by the two devices, let the corresponding detected angles be $\phi_A, \phi_B \in (-50,50)$. These angles are positive when the sound source is estimated to be on the left side of the device and negative when the source is estimated to be on the right of the device (fig. 4). We will consider the triangle that is created, with A, S and B as its vertices. The altitude of the triangle that is passing from vertex S, divides L into a and b so that a+b=L. Let the length of the altitude (in our case the distance of the audio source/person from the wall) be X_s. Then, we can formulate the following equations:

$$\tan(\theta_A - \phi_A) = \frac{X_s}{a}$$

$$\tan(\theta_B + \phi_B) = \frac{X_s}{b}$$

Since L=a+b, the final solution to the system of equations is given by:

$$X_s = \frac{\tan(\theta_A - \phi_A) \cdot \tan(\theta_B + \phi_B) \cdot L}{\tan(\theta_A - \phi_A) + \tan(\theta_B + \phi_B)}$$

$$a = \frac{X_s}{\tan(\theta_A - \phi_A)}$$

$$b = \frac{X_s}{\tan(\theta_B + \phi_B)}$$

Thus, we can calculate the precise position of the audio source in the 2-D layout of the room.

Due to the nature of the sensor and propagation of sound waves some restrictions had to be imposed in order to ensure reliable location estimation. Therefore, the sound level is calculated for a window of 1 second and the sound source angles are taken into account only when the sound level exceeds 50dB, corresponding to a quiet conversation. This technique prevents inaccurate location estimation by ignoring low level background noise. Additionally, we only calculate the person's location when the confidence for both estimated sound source angles is more than 50%. A final and apparent restriction is that there must exist a solution for the equation system and this solution should fall within the monitored space. Thus, if a sound is coming from behind the sensors, or outside the limits of the monitored space, the location cannot be estimated or it is ignored respectively. This way, noises that are generated from external sources, e.g. a car passing-by, will not affect the location estimation.

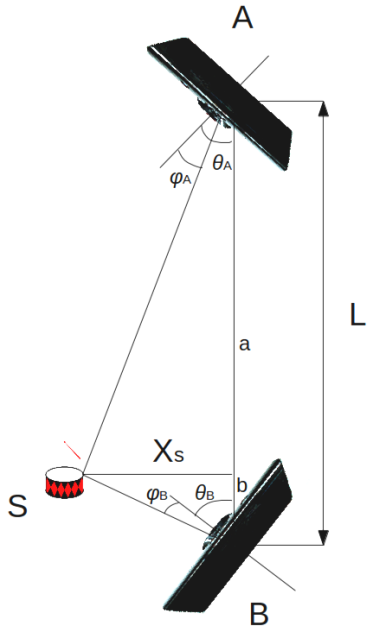

Figure 5. Configuration of the 2 Kinect devices for audio localization.

C. Automatic speech recognition module

In order to ensure natural interaction of the user with the system as well as effective emergency detection, we integrated an automatic speech recognition (ASR) module. This module is built using the hidden Markov model toolkit (HTK) [19]. The input to this module is the audio signal captured by the primary Kinect. The features extracted from the audio signal are 13 Mel frequency cepstral coefficients (MFCCs) and their first and second derivatives in order to account for speech

dynamics, creating a vector of 39 features. The models used are 3 state left-to-right HMMs modeling triphones. HTK is used for both training the models and recognizing speech. The module recognizes 11 words that comprise sentences commonly used in order to ask for help or assistance, e.g. "help me", "fire".

IV. System Operation

As mentioned earlier the two main functions of our system are person localization and emergency detection. These functions utilize information from both the skeletal tracking module and the audio localization module, but the ASR module is only used for emergency detection purposes. The following sections describe the two types of operation in more detail.

A. Person Localization

The main source of location information is the skeletal tracking module. More specifically, this module detects a person as soon as s/he enters the FOV of the sensor and tracks her/him while moving in the room. The accuracy and robustness of the tracker is exceptional due to the nature of the depth sensor, so the person is tracked while standing, walking or even sitting. We consider the location of the person as the average of the 3-D coordinates of all the tracked joints, expressed as $< \overline{X_d}, \overline{Y_d}, \overline{Z_d} >$, where:

$\overline{X_d} = \frac{1}{20}\sum_{i=1}^{20} X_{di}$ the mean distance from the sensor's plane.

$\overline{Y_d} = \frac{1}{20}\sum_{i=1}^{20} Y_{di}$ the mean deviation from the sensor's axis.

$\overline{Z_d} = \frac{1}{20}\sum_{i=1}^{20} Z_{di}$ the mean distance from the floor.

Another source of location information is the audio localization module. It should be noted that the audio localization module is capable of estimating the location of the person in 2 dimensions expressed by $< X_s$, a >, not accounting for height.

The final estimated location of the person is a result of combining the information from both modules. More specifically, when a location estimate is available from both modules, the average of each of the 2-D coordinates is calculated after proper transformation to match the 2 coordinate systems, while the third coordinate equals that of the skeletal tracking module. In the case where either of the modules does not return any coordinates, then the other module's coordinates are considered, e.g. if the person is outside the FOV of the depth sensor, only the audio localization coordinates are used. For our application, the detected activity is bound to the estimated location of the person. Therefore, if a person is standing by an appliance such as the oven or refrigerator we infer that s/he is using this particular appliance.

B. Emergency Detection

Our system is also capable of recognizing emergencies and due to the 3-D localization information it is very successful at detecting falls. In order to effectively carry out this operation, we utilize information from all three modules. In more detail, when coordinate $Z = \overline{Z_d}$ of the tracked skeleton falls below a predefined threshold (default is 2ft.), the system enters a stand-by mode. While in this mode, the system detects an emergency if any of the following conditions is met:

1. The ASR module recognizes that the person is asking for help.

2. A loud noise is recorded.

3. Z remains below the threshold and no sound is detected for a predefined period of time (default is 2 min.).

In addition, an emergency is detected even if the system is not in the stand-by mode when either of these conditions occurs:

1. The ASR module recognizes that the person is asking for help and skeletal tracking fails to locate the person.

2. The ASR module recognizes any of the predefined sentences for help repeated 3times in a 15 second window, independently of the status of the localization modules.

When any of the above 5 situations is identified, an emergency is detected and a distress signal is issued, including the person's last known location, in order to request for assistance.

V. Experimental Setup

An extensive set of evaluation experiments were conducted in order to fine-tune the parameters of the setup at our simulated assistive apartment (fig. 6). As mentioned earlier, two Kinect devices were used, mounted at the opposite sides of one of the walls, facing the entrance. The distance between the two devices was 175.5 inches. The axis perpendicular to the device points at 45 degrees towards the interior of the apartment, maximizing both the FOV and microphone coverage (fig. 7).

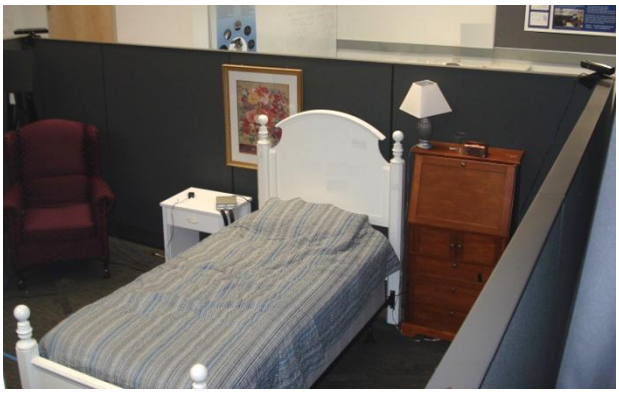

Figure 6. An aspect of our simulated assistive apartment.

All modules were installed on the same computer, although our system's implementation permits the use of separate computers for each one of the modules. For our experiments we partitioned the space in 8 different sectors, intersecting at the center of the room. The estimated location of the person was considered accurate when the coordinates fell within the boundaries of the corresponding sector. For our application, the detected activity is bound to the estimated sector.

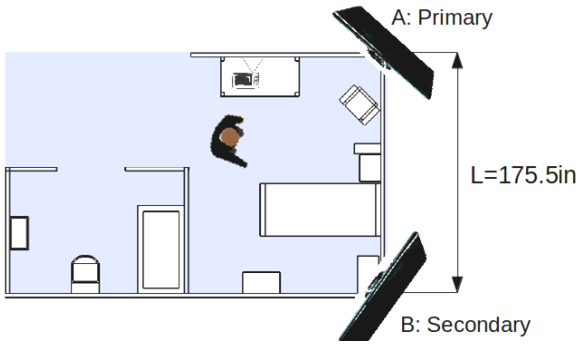

Figure 7. Simulated assistive apartment layout and placement of the Kinect devices.

Four individuals participated in our experiments, with either one or two occupying the apartment simultaneously. Subjects were asked to move in the apartment and perform every day activities. In the scenario where the space is inhabited by a single person, the location estimation of the person is always accurate as long as the person remains in the FOV of the sensor. Furthermore, in the event of 2 people being tracked, a realistic case where a caretaker or visitor is also present, the localization accuracy is 85% due to the people interacting and the resulting occlusions. In this case, person identification is not required, since we are primarily interested in the performed activities. Additional people may also be tracked but with reduced accuracy, however this is outside the scope of our system, since we aim at detecting activities and emergencies when the inhabitant is not supervised. Finally, the word accuracy of our ASR system was 94%, with low background noise levels.

VI. Conclusions

We presented a novel system capable of accurate and robust person localization and emergency detection. This system uses as input the depth sensor and microphone array of the Kinect device. Skeletal tracking and sound source localization are combined in order to estimate the position of the inhabitant. ASR is used as a natural means of interaction and in addition to the location information for emergency detection. The system was deployed in a simulated assistive environment and during the experiments conducted, it achieved both high localization and word recognition accuracy. After, confirming the effectiveness of our design we plan to extend it by utilizing depth information of additional Kinect devices for increased robustness and coverage.

Acknowledgment

This material is based upon work supported by the National Science Foundation under Grants No. NSF-CNS 1035913, NSF-CNS 0923494.

The authors would like to thank UTARI for its support.

References

[1] C. H. Kuo and R. Nevatia, "How does Person Identity Recognition Help Multi-Person Tracking?", *In Proc*. CVPR, pp. 1217-1224, 2011.

[2] D. Mitzel, E. Horbert, A. Ess and B. Leibe, "Multi-person tracking with sparse detection and continuous segmentation", *In Proc*. ECCV, pp. 397-410, 2010.

[3] M. Andriluka, S. Roth and B. Schiele, "People Tracking-by-Detection and People Detection by Tracking," In Proc. CVPR, pp. 1-8, 2008.

[4] B. Wu, R. Nevatia, "Detection and Tracking of Multiple, Partially Occluded Humans by Bayesian Combination of Edgelet Part Detectors", In Proc. IJCV, pp. 247–266, 2007.

[5] J. Krumm, S. Harris, B. Meyers, B. Brumitt, M. H. and S. Shafer, "Multi-camera multi-person tracking for easy living", In Proc. IEEE IWVS, pp. 3-10, 2000.

[6] D. Focken and R. Stiefelhagen, "Towards Vision-Based 3-D People Tracking in a Smart Room", In Proc. IEEE ICMI, pp. 400-405, 2002.

[7] A. Ess, B. Leibe, K. Schindler and L. V. Gool, "Robust Multi-Person Tracking from a Mobile Platform", In Proc. PAMI, pp. 1831–1846, 2009.

[8] V. M. Lee, T. W. Wong and C. C. Lau, "Home accidents in elderly patients presenting to an emergency department", Accident and Emergency Nursing, 7(2): 96–102, 1999.

[9] J. M. Hausdorff, D. A. Rios and H. K. Edelber, "Gait variability and fall risk in community–living older adults: a 1–year prospective study", Archives of Physical Medicine and Rehabilitation, 82(8): 1050–6, 2001.

[10] M. C. Hornbrook, V. J. Stevens, D. J. Wingfield, J. F. Hollis, M. R. Greenlick and M. G. Ory, "Preventing falls among community–dwelling older persons: results from a randomized trial", The Gerontologist, 34(1):16–23, 1994.

[11] J. A. Stevens, "Fatalities and injuries from falls among older adults", MMWR, 55(45), 2006.

[12] H. Nait-Charif and S. J. McKenna, "Activity summarisation and fall detection in a supportive home environment", In Proc. ICPR, vol.4, pp. 323- 326, 2004.

[13] C. Rougier, J. Meunier, A. St-Arnaud and J. Rousseau, "Fall Detection from Human Shape and Motion History Using Video Surveillance", In Proc. AINAW, vol.2, pp. 875-880, 2007.

[14] H. O. Alemdar, G. R. Yavuz, M. O. Ozen, Y. E. Kara, O. D. Incel, L. Akarum and C. Ersoy, "Multi-modal fall detection within the WeCare framework", In Proc. ICIPSN, pp. 436-437, 2010.

[15] "The PrimeSensor Reference Design", [online] available at: http://www.primesense.com/?p=514

[16] C. Liebe, C. Padgett, J. Chapsky, D. Wilson, K. Brown, S. Jerebets, H. Goldberg and J. Schroeder, "Spacecraft hazard avoidance utilizing structured light", In Proc. IEEE Aerospace Conference, pp.10, 2006.

[17] S. Rowe and C. Wagner, "An Introduction to the Joint Architecture for Unmanned Systems (JAUS)", Technical Report from Cybernet Systems Corporation, available at: http://www.cybernet.com

[18] "The Microsoft Kinect SDK", [online] available at: http://msdn.microsoft.com/en-us/library/hh855347.aspx

[19] S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev and P. Woodland, "The HTK Book", Cambridge Univ. Eng. Dept., Tech Rep., 2002.

Authors Profile

Georgios Galatas is pursuing his PhD in Computer Engineering at the Heracleia Human-Centered Laboratory of the Computer Science and Engineering department of the University of Texas at Arlington. He is also a research fellow of the Institute of Informatics and Telecommunications of the National Center for Scientific Research "Demokritos". He received his combined bachelor's and master's degree from the Electrical and Computer Engineering department of the University of Patras in 2008. His research interests include Computer Vision, Digital Image Processing, Automatic Speech Recognition and Person Localization. He has co-authored several peer reviewed papers published in technical conferences and journals.

Shahina Ferdous received her PhD in Computer Science from the Department of Computer Science and Engineering of the University of Texas at Arlington in December 2012. She completed her B.Sc. from the Department of CSE of the Bangladesh University of Engineering and Technology (BUET) in 2006. She worked as a Software Engineer in Therap BD Ltd. during 2007-2008, until she joined the BS to

PhD program in 2008. She joined the Heracleia Human-Centered Computing Laboratory as a Research Assistant in 2009. Her research interests include Data Mining, Database, Sensor Network and Pervasive Assistive Applications. Dr. Ferdous has co-authored several peer reviewed papers published in technical conferences and journals. She has also served as a committee member and reviewer in many conferences.

Fillia Makedon is Distinguished Professor and Department Head of Computer Science and Engineering at the University of Texas at Arlington. She received her PhD in Computer Science from Northwestern University in 1982. Between 1991-2006, she was professor of computer science at Dartmouth College where she founded and directed the Dartmouth Experimental Visualization Laboratory (DEVLAB). Prof. Makedon has received many NSF research awards in the areas of trust management, data mining, parallel computing, visualization and knowledge management. She is author of over 300 peer-reviewed research publications. She directs the Heracleia Human-Centered Laboratory that develops assistive technologies for human monitoring and smart health. She is member of several journal editorial boards and chair of the annual PETRA conference (www.petrae.org).