MULTIMODAL INTERACTION IN AMBIENT INTELLIGENCE ENVIRONMENTS USING SPEECH, LOCALIZATION AND ROBOTICS

by

GEORGIOS GALATAS

Presented to the Faculty of the Graduate School of
The University of Texas at Arlington in Partial Fulfillment
of the Requirements
for the Degree of

DOCTOR OF PHILOSOPHY

THE UNIVERSITY OF TEXAS AT ARLINGTON
December 2013

Acknowledgements

I would like to thank my supervisors, Professors Fillia Makedon and Gerasimos Potamianos for their invaluable support and guidance during my doctoral studies. I would also like to thank my committee members, Dr. Vangelis Karkaletsis and Professors Heng Huang and Vassilis Athitsos. Finally, I would like to thank NCSR Demokritos for providing me with the opportunity to participate in the joint Ph.D. program with UT Arlington. The research presented in this dissertation has been funded by the National Science Foundation through grants CNS 0923494, CNS 1035913, IIS 1041637, 1258500, IP 1338930 and the European Union through grants AVISPIRE, USEFIL.

November 14, 2013

Abstract

MULTIMODAL INTERACTION IN AMBIENT INTELLIGENCE ENVIRONMENTS USING SPEECH, LOCALIZATION AND ROBOTICS

Georgios Galatas, PhD

The University of Texas at Arlington, 2013

Supervising Professors: Fillia Makedon and Gerasimos Potamianos

An Ambient Intelligence Environment is meant to sense and respond to the presence of people, using its embedded technology. In order to effectively sense the activities and intentions of its inhabitants, such an environment needs to utilize information captured from multiple sensors and modalities. By doing so, the interaction becomes more natural as well as accurate and robust. We have focused on 3 aspects of such an environment, using speech, localization and robotics. Speech is one of the most natural forms of communication for humans. Therefore, it can be used as one of the main information sources for deriving the intentions and needs of a person. In our work, we have extended the traditional speech recognition paradigm by introducing 3 dimensional visual articulation information for recognizing spoken words. The development of our system included the capture of a novel dataset, implementation and extended testing under a variety of audio-visual noise types, demonstrating the usefulness of 3D visual information for this task. Additionally, person localization and identification is of paramount importance in a smart environment, since by knowing each person's location, her/is actions can be derived and abnormal patterns can be recognized. Our implementation conducts person identification by means of RFID. Furthermore, three types of input are combined for multi-person localization, namely, skeletal tracking, audio localization and RFID signal strength. The system was deployed and tested in our simulated assistive apartment exhibiting high accuracy. Finally, every domestic environment changes dynamically over time, creating the need for altering the position, orientation and type of sensors used within it. In our approach, we developed a framework of sensor bearing robots with the ability to relocate automatically to compensate for such a dynamic environment. Their positioning is done in such a way so as to maximize coverage. Navigation is carried out using visual information, and autonomous placement uses a decentralized algorithm.

Table of Contents

List of Illustrations

List of Tables

Chapter 1

Introduction

1.1 Problem and Motivation

Ambient Intelligence (AmI) describes an environment that is sensitive and responsive to the presence of people. An AmI environment (AmIE) improves interaction of its occupants with the technology that is embedded within its. The AmIE distributed technology facilitates everyday activities and assists people in case of need in a natural and effective way while the interconnected infrastructure remains unperceivable. The inception of such an environment happened not long before the turn of the century, with the goal to integrate a number of devices and technologies for fulfilling this goal, leading to the first attempt to create such an environment by MIT [1]. Precursors to this concept were human-centered [2], ubiquitous/pervasive computing [3] and context-awareness.

Human-centered computing is the field of study according to which a human is affected by computational processes with the intention to improve her/is quality of life [4].

The main axes of emphasis for such systems are:

- Access to resources through interacting with a seamless computing infrastructure that allows for changes to the physical or virtual environment. These changes should lead to the improvement of the quality of life of the person while being unintrusive and conserving her/is privacy
- Detection of events and episodes (series of events [5]). This should allow for the recognition of abnormalities, as well as emergency detection and risk prevention.
- The ability to recognize and adapt to changes of the environment in an on-line manner, by utilizing monitoring information from static and mobile sensors.

In other words, the infrastructure must be: 1) embedded (integrated and networked devices), 2) context aware (able to identify people and events), 3) personalized: (tailored to the user's needs), 4) adaptive (change in response to the user) and 5) anticipatory [6].

An environment can be characterized as context-aware when it has the ability to utilize sensors to collect and interpret information that enables sensing an entity or a situation [7], [8] . This is a relatively recent concept, and it is based primarily on the crucial research fields of location awareness and activity recognition [9]. These problems can be approached using different modeling techniques [10].

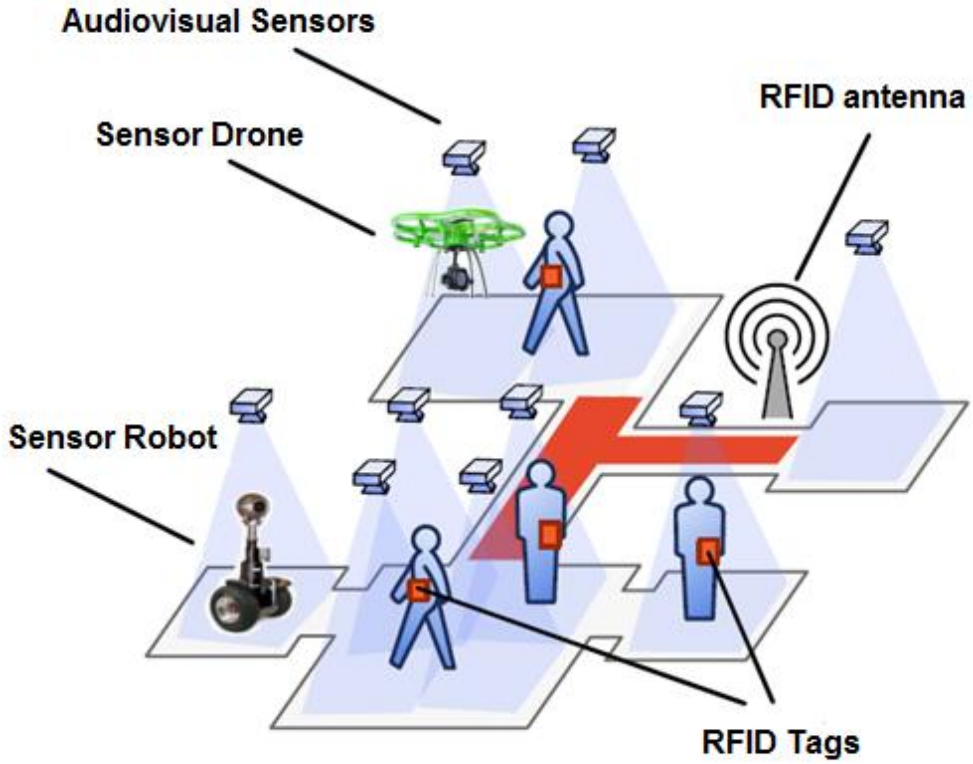


Figure 1-1: Our vision for a domestic ambient intelligence environment, incorporating audiovisual sensors, RFID and robotic platforms.

AmI applications face serious challenges concerning data collection, modeling and analysis. The main difficulty springs from the multimodal nature of the data captured from the diverse human activities. Traditionally, the analysis of the different data has been based on numerous methods and techniques, increasing the difficulty of conducting, assessing and comparing different experiments and systems. Therefore, there is a need to develop appropriate tools that will enable collecting and correlating the multitude of human activity data that can be captured in an AmIE by a number of sensors. Therefore, the main focus of this work is to develop tools and methods towards a smart system that makes communication and interaction between humans and AmIEs easier, seamless and intuitive, while enhancing context-awareness.

1.2 Approach and Contributions

One of the most natural forms of communication for humans is speech. Therefore, a decisive step for improving interaction in an AmI environment is through Automatic Speech Recognition (ASR). In our work, we have developed a novel multimodal ASR system that in addition to audio, utilizes 3-D visual lip movement information captured by the Microsoft Kinect sensor in two languages. Our system goes beyond the traditional audio-visual ASR (AVASR) paradigm by being the first to incorporate depth information from the Kinect. This novel system utilizes 3 streams of information, namely audio, planar video and depth information, and is described in Chapter 2. Initially, we present the methodology followed to collect a connected digit database using multiple sensors. This database contains two parts, one in English and one in Greek, with the former comprised of 4500 digits uttered by 15 speakers and the latter, 2200 digits by 6 speakers. At the time captured, this was the only database of its kind, utilizing a structured light sensor for AVASR, and the only AV digit corpus for speech recognition in Greek. The implementation of the system incorporated the use of the Viola Jones detector as a visual front-end, followed by the use of the extraction of appearance based features and a two stage LDA for feature selection. Decision fusion Hidden Markov Models were used for data fusion of the audio, video and depth streams. We conducted a large number of experiments in both languages. The methodology followed deviates from the traditional paradigm by incorporating not only audio noise, but also four different types of visual noise to our data, in order to test the system's robustness as well as the potential of 3D lip information to AVASR. Our system exhibited high word recognition accuracy (99.02%) when utilizing all three streams. In addition, depth information was found to increase recognition performance especially for low to medium audio SNR values. Furthermore, the use of the visual modality containing depth information in conjunction with decision fusion resulted in an average increase of 22.1% when compared to an audio-only recognizer. Thus, our experiments proved the usefulness of 3D visual information from the Kinect for the AVASR task in both languages and the effectiveness of our system's design.

In addition to the communication aspect of an AmI environment, accurate and robust location-awareness is of paramount importance for context-awareness. The two main facets of location-awareness are person identification and localization. In Chapter 3 we present the development of a novel system for simultaneous multimodal, multi-person identification and localization. Our unintrusive system uses three different modalities in order to identify and track multiple people, namely RFID received signal strength, skeletal tracking from depth images and sound source localization. The data is captured by 2 Antennas and 2 Kinect sensors respectively, deployed at the corners of a simulated apartment to continuously track and identify its occupants, thus enabling activity monitoring. More specifically, skeletal tracking is carried out using the Kinect sensor's 3D depth images and sound source localization is conducted utilizing microphone arrays of 2 such sensors, to deduce accurate location information. At the same time, the video information is not captured, making this approach less intrusive than using video cameras. RFID is used mainly for discerning between users and also for providing a rough estimate of their location by means of the signal strength indicator, enabling mapping the location information from the Kinect sensors to the identification events of the RFID. Our system was evaluated in a real world scenario involving the simultaneous real-time identification and location estimation of 4 individuals. Our goal was to identify and map the location of each person in a simulated apartment setting at a detailed level that would allow inference of conducted activities. During our experiments, it demonstrated high robustness and multi-person localization accuracy exceeding 90% using the Kinect, which constituted the most accurate source for person location information. The promising results attained, showed the great prospect of using the RFID and Kinect sensors jointly to solve the simultaneous identification and localization problem.

In order to satisfy the requirement for adaptability in AmI, the entire monitoring infrastructure that we have described so far should be able to change dynamically over time to account for the changes of the environment. This fact creates the need to alter the position, orientation and type of sensors used. In Chapter 4, we present the development of a novel system that optimally positions a number of sensor bearing robots with the ability to relocate them automatically to compensate for such a dynamic environment allowing their use for event recognition and guidance. The existence of more than one robot can ensure more effective monitoring capabilities and responsiveness in case of need by the user. In such a case, the problem of positioning the robots in the apartment is of utmost importance, since a fully automatic and efficient algorithm must be used to ensure effective and optimal coverage of the monitored space by the robots' sensors, while taking into account both the layout of the space and the characteristics of the sensors. Therefore, we utilized a tool for sensor placement and system monitoring that allows for designing the environment layout, defining crucial areas and updating the position of the robots dynamically. Placement of the robots is done by means of the Extended Max Sum Decentralized Coordination (EMSDC) algorithm, in such a way so as to maximize coverage and account for changes in the environment and detected events. The operation pipeline for the system would involve the definition of the apartment layout and the areas of high importance, creating a "critical area map". After these regions, as well as the number of robots available have been defined, the tool uses the EMSDC algorithm to define the number of the robots and position of each one in the different areas of the apartment. Our system was based on the prototype assistive-guide robot eyeDog, developed initially to provide the visually impaired with autonomous vision-based navigation and laser-based obstacle avoidance capabilities. The components of the robot are the Create robotic platform (iRobot), a net-book, an on-board USB webcam and a LIDAR unit. The camera is used as the primary sensor for the navigation task; by means of vanishing point estimation. The controller module steers the robot utilizing a PID controller, which guarantees that the robot moves following the direction of the path. While moving, the robot uses the LIDAR for obstacle avoidance. Most related research has focused on vehicle traffic instead of indoors environments, and similar implementations utilize and RFID technology for navigation purposes. The novelty of this implementation is the use of RANdom SAmple Consensus (RANSAC) with adaptive thresholds to estimate the vanishing point of the visual scene as well as the fusion with the laser data in order to avoid obstacles while navigating. During the evaluation phase, the eyeDog robot proved to successfully navigate on a given path and avoid obstacles. Furthermore, our novel framework for optimal assistive robot placement enabled robots incorporating a variety of different sensors to navigate and effectively monitor an apartment given its layout, the importance of different rooms and the user’s preferences.

Chapter 2

Multimodal Automatic Speech Recognition

2.1 Introduction and Related Work

Human speech constitutes one of the most natural forms of communication; therefore, Automatic Speech Recognition (ASR) has been drawing the ever-growing interest of the research community. The incorporation of visual information for ASR has been utilized as a means for recognition robustness, enabling natural, speech-based human-computer interaction. The benefits that spring from this approach are straightforward; weaknesses of one modality are offset by the strengths of another, resulting in better performance. In this chapter, we investigate the use of 3D visual information captured by the MS Kinect for the task of audio visual automatic speech recognition in two languages. We present our novel system that utilizes information from three streams, namely audio, planar video and depth information. This system employs appearance based visual features and LDA for feature selection, as well as decision fusion HMMs for statistical ASR. We also demonstrate the data capturing methodology used to collect our database using multiple sensors. We have conducted extensive experiments on our system in both English and Greek. Our methodology deviates from the traditional experimental paradigm by incorporating a variety of audio-visual noises in order to test the system's robustness. Our results show that depth information from the Kinect benefits lip reading performance and that 3D visual information increases word accuracy considerably, in comparison to conventional audio-only ASR in both languages.

The joint use of acoustic features and visual information extracted from the speaker's mouth, as in the Audio-visual speech recognition (AVASR) paradigm, has been investigated in the literature and found to increase ASR accuracy and robustness in the presence of acoustic noise [1] - [13].Typically, the incorporated visual information is extracted from planar video data of the speaker’s face, captured in the visible spectrum and results in considerable recognition improvements, when combined with noisy audio information within a two-stream classifier fusion framework. Nevertheless, in this traditional paradigm useful 3D visual speech articulation information from the speaker is not utilized. Few only attempts have been made that deviate from this paradigm, employing multiple cameras to capture the speaker's face, with an increase though in hardware cost and software complexity.

A crucial requirement for designing an ASR system is the availability of appropriate corpora that allow investigating the various aspects of the research problem of AVASR. A number of audio-visual datasets exist in the literature. One of the most popular ones is CUAVE, which is a database of isolated and connected digits (0-9) in English uttered by 36 subjects [14]. Other databases include Tulips1, a 12-subject database in English of digits 1-4 [15], DAVID, a database of 31 speakers which includes digits, alphabets, and vowel-consonant-vowel (VCV) utterances [16] and XM2VTSDB including 295 speakers uttering 2 sequences of digits and a sentence each [17]. There are also some databases where stereo-cameras were used to capture the 3D information of the speaker’s face. Such a database is AVOZES which includes digits 0-9, VCV, CVC utterances and 3 sentences by 20 speakers in Australian English [18], later used in [19] for 3D lip tracking. In [20], a Bumblebee stereo-camera is used for collecting the WAPUSK20 database that consists of stereoscopic video and 4-channel audio of 20 speakers uttering 100 sentences in English. Finally, the AV@CAR database is a multimodal corpus of in-vehicle videos in Spanish, which provides 6 pictures of each subject in order to reconstruct the 3D textured mesh of the speaker’s face [21]. In our work, we have followed an alternative approach, aiming to capture 3D visual speech information. Therefore, we move beyond the traditional visual stream paradigm, by incorporating facial depth data, captured by a novel multimodal device, the Microsoft Kinect [26], [27]. We combine this data with the audio and video streams, thus employing facial depth information. Our database is collected using multiple audio and visual sensors. Among them, of particular interest is the use of a novel, popular and affordable device, the Microsoft Kinect, that operates based on the structured light method [22], aiming at capturing 3D information of the speaker's face. No other such database that uses the structured light technique exists for AVASR. In addition to the depth video, the corpus also contains traditional color video, from both the Kinect and an HD digital video-camera. Furthermore, the first AV digit corpus in Greek was captured as part of our database.

Another important aspect of our effort is related to extracting informative features from the visual and depth streams. For this purpose, various feature selection and transformation techniques have been adopted in the literature. Various schemes have been proposed, such as, the use of genetic algorithms for feature selection and principal component analysis for feature transformation [23]. Appearance based features, obtained from the discrete cosine transform (DCT) of the mouth region-of-interest (ROI), have been employed in our approach. A straightforward feature selection method of the resulting DCT coefficients is the use of feature energy as a measure of information content [24]. According to this technique, features with higher energy values over time are more informative, and thus their selection based on energy sorting can be effective. The process is further facilitated by the use of linear discriminant analysis (LDA) that has been observed to benefit automatic speech recognition performance [25]. In our work, LDA is applied both within each frame and across temporally adjacent feature frames to capture dynamic speech information. Furthermore, this two-stage LDA is also applied on the depth data, after appropriately mapping the tracked ROI from the traditional video to the depth data stream.

The robustness of our system was investigated in an extended number of experiments not only under the mainstream audio noise conditions, but also under much less studied visual noise conditions [28], exploring the potential of 3D lip reading information to AVASR. In particular, we consider four types of visual noise that can be encountered in typical application scenarios of AVASR, such as degraded contrast and brightness conditions, as well as Gaussian and block noise resulting from sensor or transmission channel failures [29]. We also report experimental results on the two different languages of our database, English and Greek. The resulting multisensory, multimodal AVASR system has been demonstrated to yield superior performance when using the additional modalities for the ASR task across both languages.

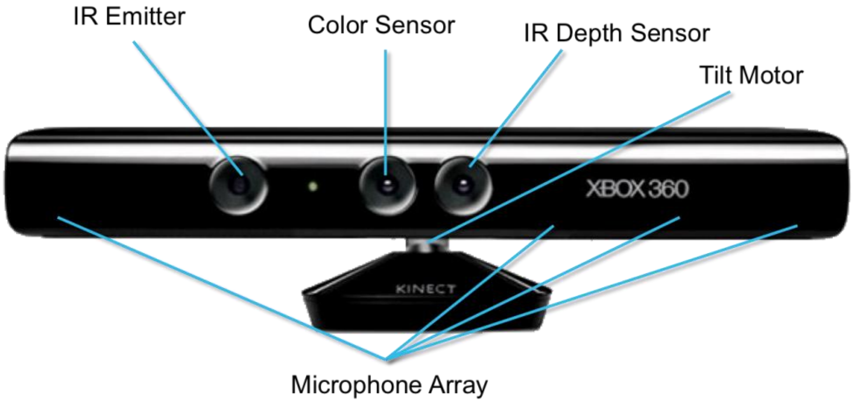


Figure 2-1 The Microsoft Kinect device indicating the location of the various sensors, IR emitter and tilt motor

2.2 Data Acquisition Hardware and Setup

Multiple devices were used during the data collection process in order to capture the different types of data. More specifically we used a High Definition video-camera, an external voice recorder, and the MS Kinect.

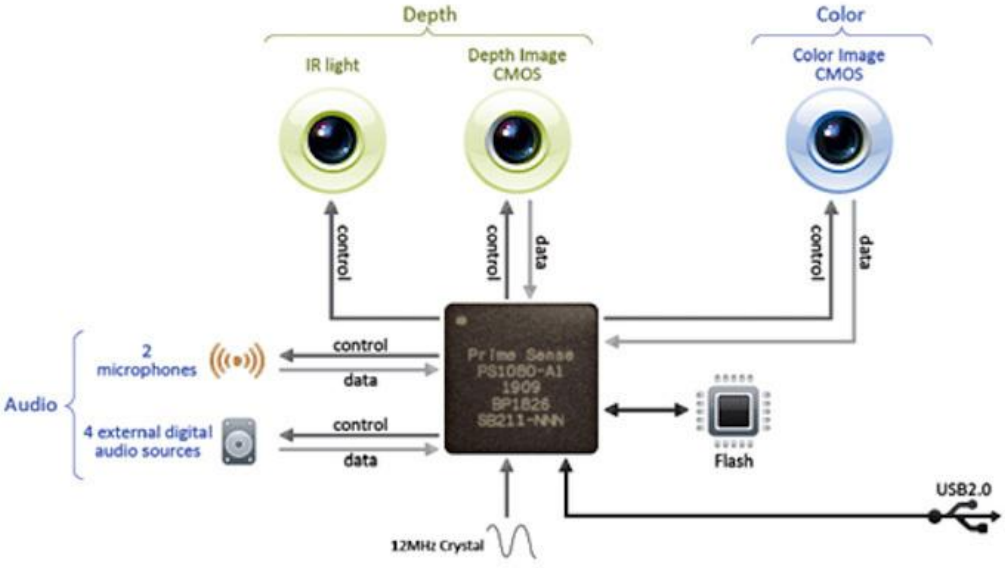


Figure 2-2: The PrimeSensor device reference design, showing the distance measurement configuration, color image sensor and audio input [30].

2.2.1 Microsoft Kinect

The Microsoft Kinect (shown in Figure 2-1) is a novel device developed mainly for gesture recognition. It is based on the PrimeSensor device design [30] shown in Figure 2-2 and in addition to VGA resolution video (640x480 pixels), it can also capture depth images of the same resolution. In order to capture depth images a laser, an IR camera, and the structured light methodology are used (Figure 2-3). This technique works as follows; first, the laser beam passes through a grating, where it is split into many 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 [22]. The effective range of the depth camera hardware of the first generation Kinect is 2.3-20ft. Thus a minimum distance between the sensor and the speaker must be maintained at all times, since too short of a distance can prohibit distance measurements from being captured. In our experimental setup, the video and depth streams were 640x480 pixel, 24-bit RGB at 20fps and 640x480 pixel, 11-bit at 20 fps respectively. Since at the time of the data collection there existed no interface for the Kinect's microphones through USB, we used a high quality external voice recorder for capturing audio.

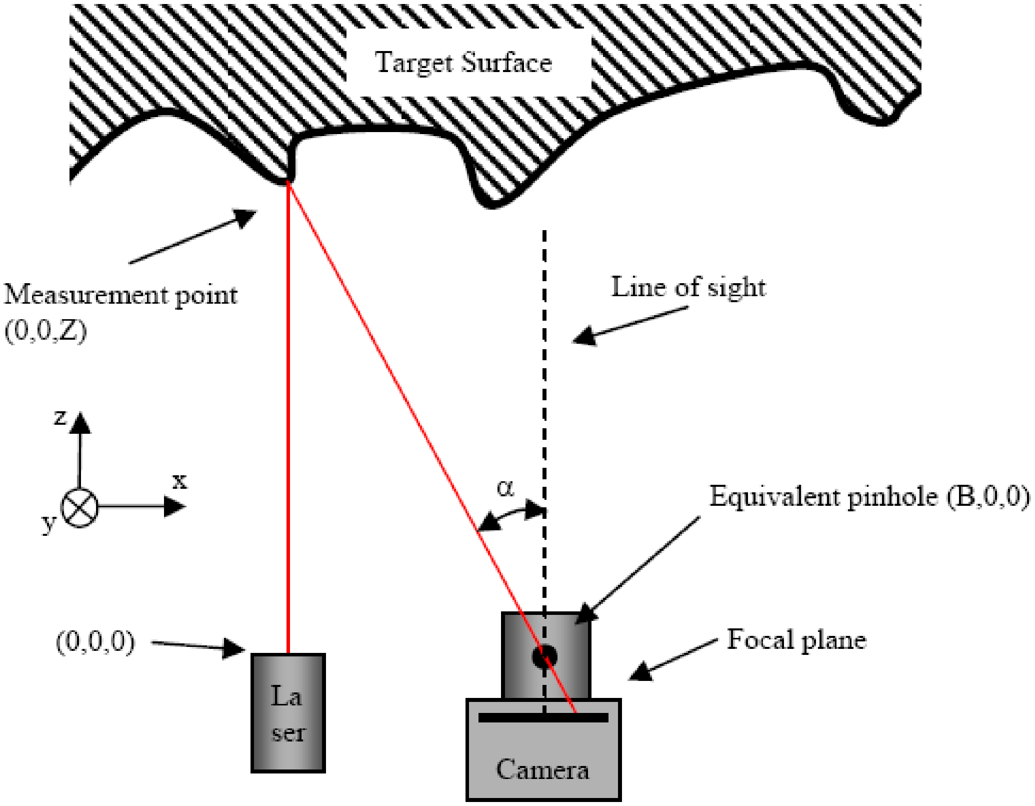


Figure 2-3: The structured light methodology for distance estimation through triangulation [22].

2.2.2 Other Devices Used

In addition to the Kinect, we used a Zoom H4 voice recorder and a Cannon Vixia HF100 HD camera. The voice recorder incorporated a pseudo X/Y condenser microphone setup which exhibits nearly uniform directionality and relatively flat frequency response. Its specifications allowed capturing 24-bit 96Khz sound in 2 external tracks but for our experiments we captured sound in stereo 16-bit 44.1Khz PCM format, which is more than adequate for voice signals. The digital video-camera captured HD 1080p video at 29.97 fps in MPEG TS format.

2.2.3 Setup

All three devices were used to collect audio and video data. The Kinect sensor has two cameras, one IR camera for measuring the distance and an RGB camera, that are 1 inch apart. In order to minimize lateral disparity, we rotated the whole device by 90 degrees. Furthermore a calibration pattern (chessboard) was captured from various angles which could be used for calculating precisely the correspondences between the RGB and the depth image. The Kinect was placed at approximately 2.95 ft. from the speaker's mouth, because of the range restrictions of the depth sensor. The HD camera was placed as near as possible to the Kinect’s axis, in order to capture images as much as possible with the same angle as the Kinect. The voice recorder was placed between the other two devices facing the speaker. A monitor behind the devices prompted the speaker to utter specific number sequences. The position of the monitor ensured that the subject was facing the cameras while speaking. The whole capturing and prompting process was automated and controlled using appropriate Perl scripts. The devices and setup can be seen in Figure 2-4.

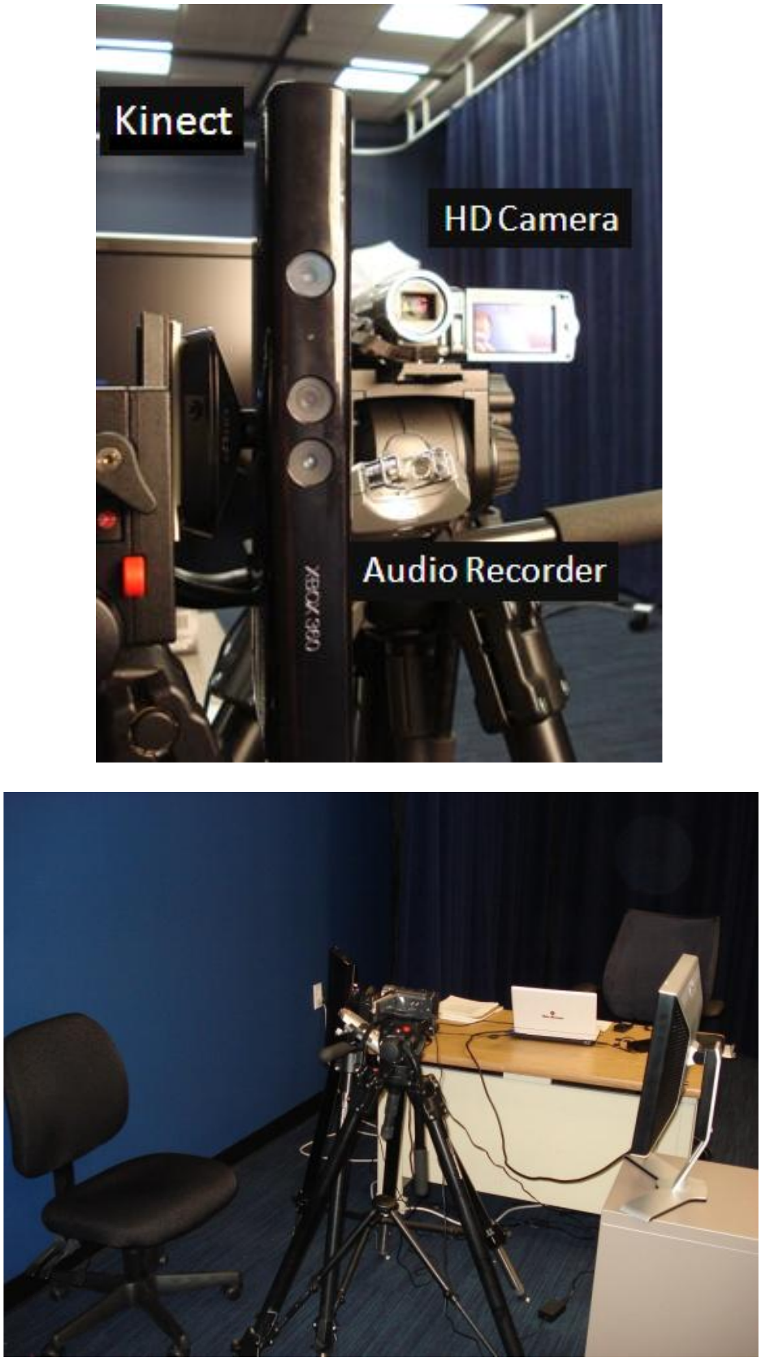


Figure 2-4: The data collection process displaying the devices (top) and configuration (bottom).

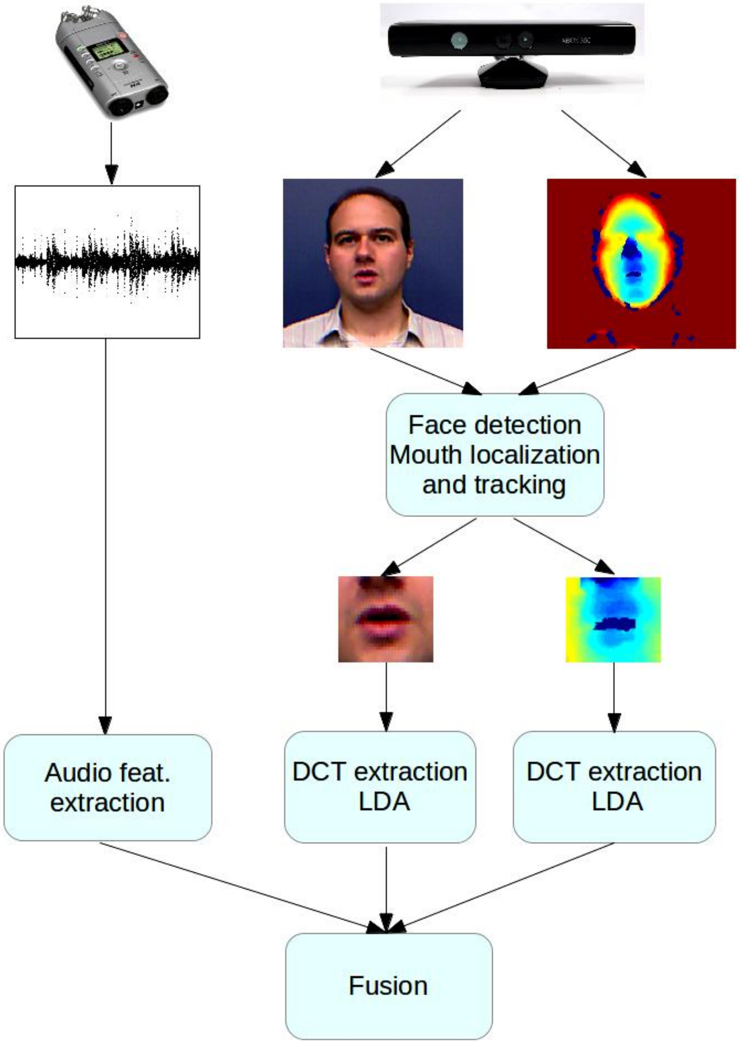


Figure 2-5: An overview of the architecture of our multimodal AVASR system with details on the visual front-end and feature extraction.

2.3 Theory and System Architecture

Our system is composed of the following modules: 1) the visual front-end, implementing the region-of-interest detection, 2) the feature extraction and transformation module and 3) the statistical ASR module for model training and testing on features fused across all data streams. A system overview is depicted in Figure 2-5. The modules are described in more detail in the following sections.

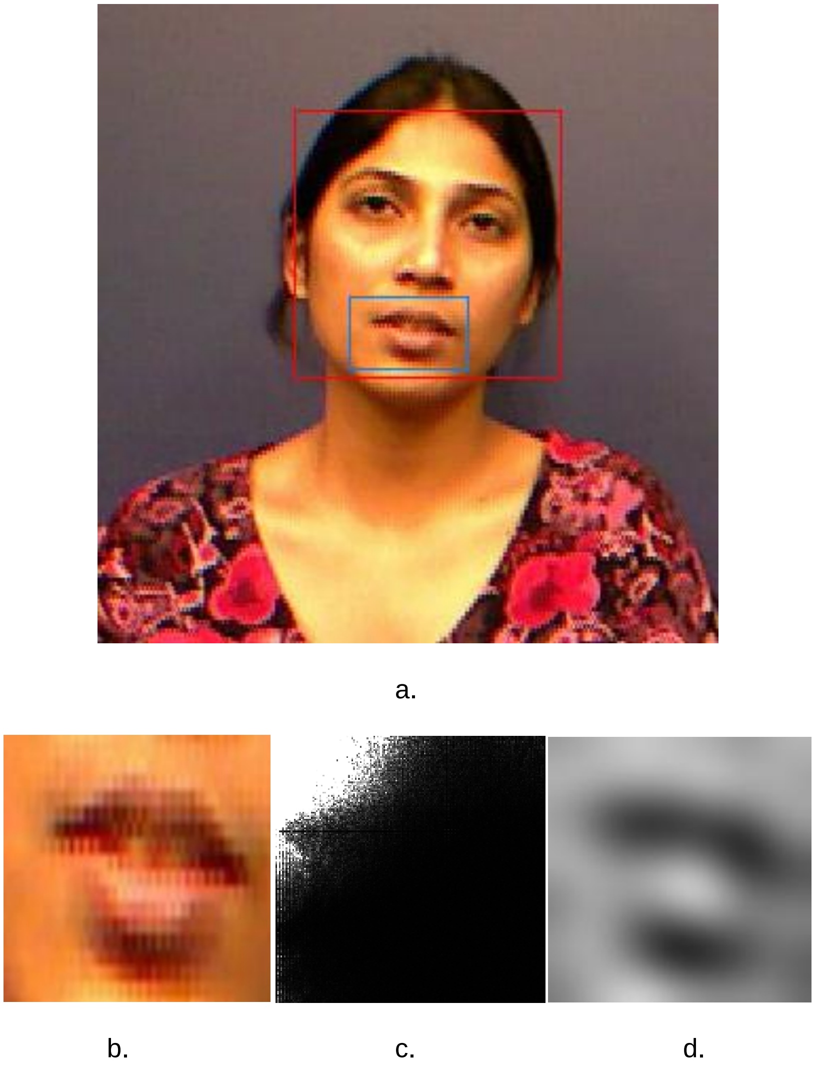


Figure 2-6: Visual front-end and feature extraction. Examples for a. tracking, b. extracted ROI, c. 2-D DCT image and d. inverse DCT image.

2.3.1 Visual Front-End

The visual front-end can be divided in two parts: i) Face detection and mouth localization, and ii) Visual feature extraction.

The Viola-Jones detector is used in order to achieve robust and real-time face detection. The Viola-Jones algorithm employs AdaBoost, a binary classifier that uses cascades of weak classifiers to boost its performance [31], [32]. This detector is utilized twice, in our implementation. Initially, the face in each frame of the video sequence is detected and subsequently the mouth portion of the face is detected. By conducting this nested detection for the mouth, the number of false positives in the image is decreased, while preserving the performance at a high level. In addition, the coordinates of the bounding box of the mouth are filtered, to make sure that false detections and abrupt movements do not hamper the mouth tracking process. This is achieved by finding the median of the coordinates of the bounding box in a 10 frame window. The resulting coordinates were also used for locating the mouth region in the depth images, by adjusting the coordinates according to the disparity of the two sensors. The final ROI is obtained by resizing the respective mouth bounding box to 64x64 pixels. An example of this process is depicted in Figure 2-6 a and b.

2.3.2 Feature Extraction

The next step after the ROI extraction is to express useful information of the lip movement with appropriate features, adequately capturing the speech information present in the lip movements. Such features can either focus on the contour of the mouth and face (shape-based) [33], or extract information from the whole mouth region (appearance-based) [28]. We opted for the latter, using the coefficients of the upper-left corner of the Discrete Cosine Transform (DCT) from each video and depth frame [24]. We considered the coefficients in the upper-left corner because they have higher energy values and thus capture more lip movement information. In addition, we take into account only the even coefficients of the DCT, similarly to [34]. This compensates for variations in head pose. The number of coefficients we extracted with the aforementioned methodology was 45 for every frame. The resulting coefficients of the ROI as well as the inverse transform are shown in Figure 2-6c and d respectively.

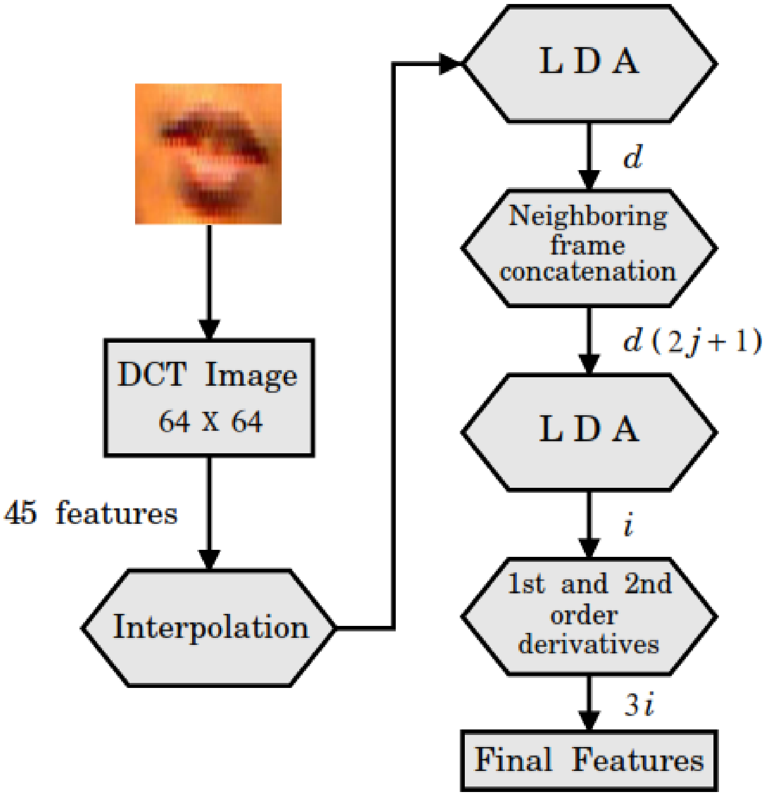


Figure 2-7: The feature extraction and selection pipeline utilizing a two-stage LDA used in our system for the visual and depth data streams.

A two-stage LDA based approach depicted in Figure 2-7 was implemented in order to improve feature selection, similarly to [25]. In more detail, at the first stage, we applied LDA on the 45 features of each frame ("intra-frame") selecting d<45 features with the highest eigenvalues. At the second stage, we concatenated j neighboring feature vectors at each side of the vector of the current frame, in order to capture dynamic visual speech information. We then applied LDA ("inter-frame") to the concatenated vector of dimension (2j+1)d, selecting a smaller number of features i with the highest eigenvalues.

Finally, we calculated their first and second order temporal derivatives, appending them to the feature vector, thus yielding features of dimensionality 3i. For the planar visual data stream, values d=10, j=3 and i=10 were chosen, whereas for the depth data stream values d=15, j=6 and i=10 were preferred. In both cases, the final features were of dimension 3i=30.

The Hidden Markov Model Toolkit (HTK) [35] was used for the audio feature extraction, in order to extract the well known Mel-frequency cepstrum coefficients (MFCC) as features, along with their first and second derivatives on windowed speech segments of 25 ms duration and 10 ms overlap. The resulting feature vector comprises of 39 elements. As a final step we need to ensure identical feature extraction rates. However, the visual feature rate is the same as the frame rate of the video, namely 20 Hz, whereas the audio feature rate is 100Hz. Therefore, the visual features are interpolated temporally so as to reach the same frame rate as the audio features.

2.3.3 Statistical ASR

Hidden Markov models (HMMs) are broadly used in ASR applications for modeling speech. The Baum-Welch algorithm is used for training the models and the Viterbi algorithm for recognition. In our experiments, we compared the performance of two types of models, baseline single-stream HMMs (i.e. feature fusion) and state-synchronous multi-stream HMMs (two- and tri-stream HMMs, i.e. decision fusion). This type of model realizes a decision-fusion approach, by computing the state emission (class conditional) probability as a product of the observation likelihoods of every stream, raised to a specific exponent λ, as shown in the following equation:

$$\Pr[o_t^{AVD} \mid c] = \prod_{s \in \{A,V,D\}} [\sum_{k=1}^{Ksc} \omega_{sck} N_{d_s} (o_t^{(s)}; m_{sck}, s_{sck})]^{\lambda_{sct}}$$

This exponent is bound to the reliability of the stream itself and defines the contribution of each stream. o_t^{AVD} denotes the tri-modal observation vector $o_t^{AVD} = \{o_t^A, o_t^V, o_t^D\}$, s is one of the three streams, c denotes the HMM state and t is the time (frame) of the utterance.

Thirty context-dependent phonetic models (triphones) were trained in total, each having three emitting states in a left-to-right topology (Figure 2-8) and four Gaussian mixtures per observation stream and state. HTK patched with the HTS software [36] were used for training and testing and a free grammar was used at decoding.

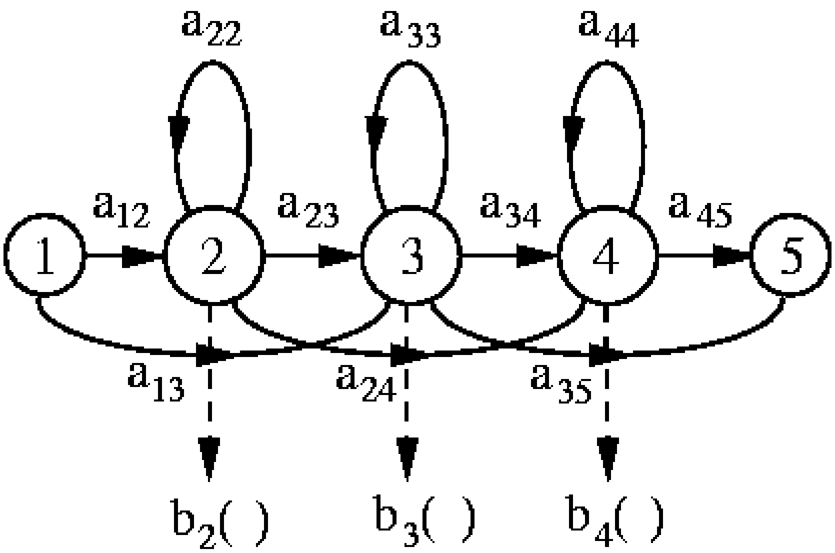


Figure 2-8: An example left-to-right hidden Markov model with 3 emitting states and 2 non-emitting (1st and 5th) [35].

2.4 The BAVCD Database

The BAVCD database was shot at the Vision Capture and Human Tracking Laboratory (now MoCap) of the Computer Science and Engineering Department at the University of Texas at Arlington. This lab was characterized by three main benefits for shooting in: 1) Background noise levels were non-existent throughout the experiment, 2)

Illumination was controlled and set to 220 Lux, 3) The background was solid blue, allowing easy detection and tracking of the person's head, as well as enabling the projection of different backgrounds using chroma-keying for face detection experiments. We collected a connected digit corpus for digits 1-10 in both English and Greek. The speakers were therefore asked to read in a continuous manner random 10-digit numbers that were displayed on a monitor by an automated script. During each session, the speaker was asked to read five 10-digit sequences, repeating the process 6 times (totally 30 ten-digit sequences). In the English corpus, both pronunciations "zero" and "oh" for digit "zero" were used with equal probability (vocabulary size 11) while the Greek vocabulary size was 10. Each speaker was given 5 seconds per sequence and the recording time for the five sequences was 35 seconds. At the beginning of each recording session the speaker was asked to clap her/is hands to facilitate synchronization of the various streams.

A number of challenges were faced and resolved when using the Kinect for capturing video. No official software or drivers existed at the time of the experiments for interfacing with the device, thus solutions developed by the open-source community had to be used. The libfreenect drivers and libraries were employed in order to capture both RGB and depth images, but no support for capturing sound existed. Furthermore, the actual framerate of the streams was 20 fps, despite the fact that according to the official specifications 30fps should be feasible. In addition, although the PrimeSensor has provision for hardware synchronization of the streams, the Kinect has no such hardware. As a result the two streams are not accurately synchronized and their corresponding frame rates fluctuate around 20 fps.

By recording the capturing timestamp of each frame, synchronization across the Kinect video streams was made possible by considering one of the streams as reference

(RGB) and normalizing the framerate. More specifically, the normalization process involves separating the stream in 10-frame segments and, by using the timestamps, calculating the framerate of every segment. Then, for every segment, if the framerate is higher than 20, frames are dropped and if it is lower than 20, nearest-neighbor interpolation is used to duplicate frames. In most cases either none or one frame needed to be either dropped or interpolated thus not harming the smoothness of the video. This also allowed for better synchronization with the audio stream as well as more precise segmentation of the sequences in each recording. Lastly, the frames of the reference stream (RGB) were matched to the frames of the other stream (depth) by finding the temporal distance of the two. In the future Dynamic Time Warping could be used for better matching of the two streams, but the results using our method are considered satisfactory.

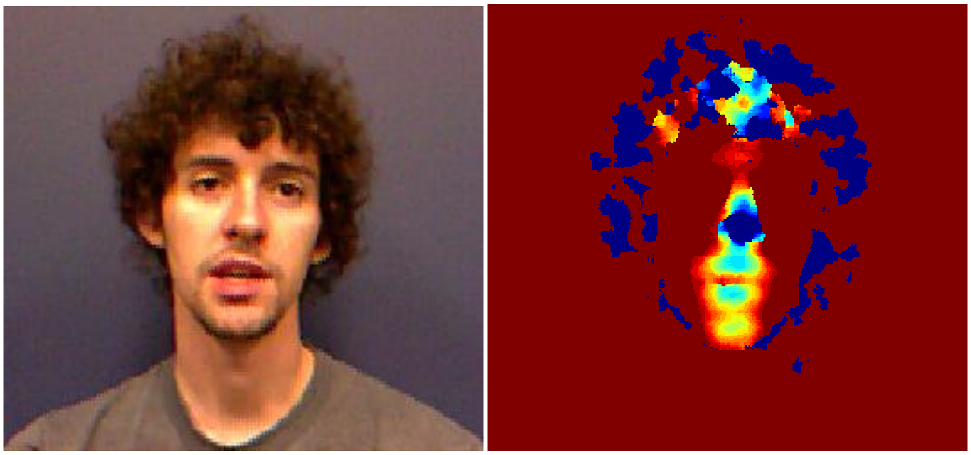


Figure 2-9: Example of a collected color video frame (left) and the corresponding depth frame (right).

Due to the fact that many different sensors were used, a clap was captured by all sensors at the beginning of the shooting, used to segment the data streams and mainly synchronize the Kinect video streams with the audio stream. An alternative approach for synchronizing the HD camera streams is possible, by employing cross-correlation and using the Zoom-H4 recorded audio of the audio streams recorded by the two devices.

2.4.1 The Data

The database is comprised of a set of connected digit utterances in two languages, English and Greek. The total number of speakers was 15, 6 of which were native Greek speakers. The English part contains approximately 4.500 connected digits from all 15 subjects and the Greek part contains approximately 2200 connected digits. The data includes the RGB and depth frames captured by the Kinect, as well as the timestamps for each frame, the corresponding audio segments captured by the audio recorder and the HD videos from the camera. All the frame sequences, audio and video files are segmented at the 10-digit sequence level. The images captured by the Kinect are saved as PNG format files because of its good compression rate and lossless nature. RGB frames were saved as ordinary three channel 24-bit PNG files, while depth frames are single channel with 11-bit accuracy and therefore they were saved as 16-bit grayscale PNG files. Since ordinary monitors cannot display images higher than 8-bit per channel, in order to display a depth frame either compression of the dynamic range of the image or, since we are interested in the mouth region, thresholding the intensity values displayed taking into account the intensities of the ROI and expanding this narrow dynamic range to 256 values (8-bit) is needed. Such an example is shown in Figure 2-9 after color mapping for visualization purposes.

2.5 Experimental Setup

In this section, we present the experimental setup devised to evaluate the performance of our system. More specifically, we conducted extensive experiments in two languages, English and Greek, under the influence of audio noise [37]. In addition, we carried out experiments using degraded video from both parts of our database. All three streams of the BAVCD database, audio, planar video and depth, were utilized. Training was conducted on clean speech, simulating the mismatch between the training and testing conditions. As already mentioned, all data utterances were connected digit sequences and the vocabulary size was 10 for Greek and 11 for English, since both “zero” and “oh” were used for digit “0”.

In the following subsections we analyze the details for each set of experiments and present the corresponding results.

2.5.1 Noise Types

The audio samples were corrupted with additive babble noise from the NOISEX-92 database [38] at several signal-to-noise ratios (SNRs), in order to test the system for robustness under the influence of audio noise.

Additionally, the performance of our three-stream AVASR system was studied under the presence of visual noise, thus moving beyond the traditional AVASR paradigm where recognition is evaluated under noise in the audio channel alone. In our preliminary work [39] we performed such a study, but on a traditional, two-stream AVASR system without depth information, studying the effect of three types of visual degradations. Here we consider 4 types of visual noise, namely Gaussian noise, block noise, reduced brightness and altered visual contrast. These degradations can be viewed as the result of faulty transmission channel or camera or could be encountered in low-lighting conditions. Since the depth sensor operates independently of lighting conditions, such noise types do not affect it, and hence its use is expected to become of increased practical significance.

The effects of the visual degradations considered to the ROI are depicted in Figure 2-10. Below, we explain in detail how they are used in our experiments.

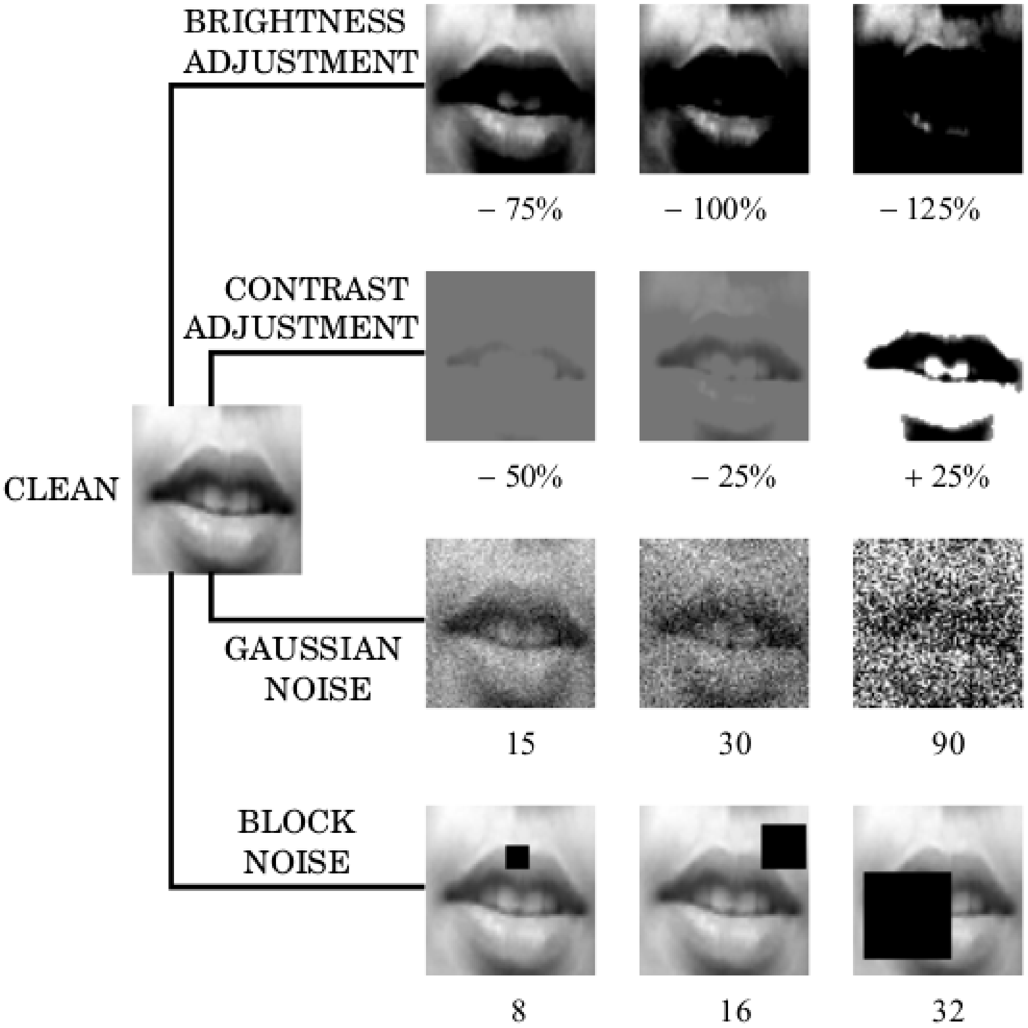


Figure 2-10: The various types of visual noise applied on the ROI at different levels and their effect.

2.5.1.1 Reduced Brightness

Reduced brightness conditions are simulated using the following algorithm:

i. The mean intensity of the ROI pixels is calculated.

ii. A percentage of variation is set.

iii. An intensity variation value is calculated by multiplying the percentage with the mean.

iv. This value is added to the intensity of every ROI pixel, thus altering its intensity.

v. The new intensity value of every ROI pixel is limited between 0 and 255.

This procedure alters non-linearly the intensity of every pixel by a constant value, but allows comparison of the amount of variation with the mean value of all pixels.

2.5.1.2 Altered Contrast

In the case of contrast, we experimented with both overly high and low values, since both can have a negative effect on image quality. For simplicity we describe only the method used for lowering contrast:

i. The mean intensity of the ROI pixels is calculated.

ii. A percentage of variation is set.

iii. An intensity variation value is calculated by multiplying the percentage with the mean.

iv. For every ROI pixel with intensity lower than the mean, this value is added, with the provision that the resulting intensity does not exceed the mean.

v. For every ROI pixel with intensity higher than the mean, this value is subtracted, with the provision that the resulting intensity is not lower than the mean.

The mean intensity value is used not only as a reference point, but also as a pivot point to vary intensity. For increasing contrast, we follow the same steps, but reverse the addition and subtraction conditions and restrict the resulting intensities within [0,255], similarly to the case of reduced brightness.

2.5.1.3 Block Noise and Gaussian Noise

Block" noise and additive Gaussian noise were the other two types of noise considered. The former consists of a black square block of predefined size that appears randomly inside the ROI at every frame causing loss of lip movement information. The sizes of the block that we considered were 8x8, 16x16 and 32x32 pixels. The latter is the well known additive Gaussian noise of zero mean and a variable standard deviation of 15, 30 and 90.

2.6 Results and Discussion

In this section, the results acquired from our experiments are presented and compared to baseline lip-reading performance without depth information and traditional audio-only ASR performance.

2.6.1 Audio Noise

2.6.1.1 English

Initially, the data from 14 subjects of the English part of the database was used in a random 2/3, 1/3 split for training and testing, respectively, to conduct multi-speaker ASR experiments.

The effects of the use of depth information in combination with planar video, as well as the positive effects of feature transformation using the two-stage LDA are presented in the first set of experiments. Table 2-1 clearly depicts the performance improvement resulting from the use of the two-stage LDA not only when compared to the energy based feature selection method (by 18.13% relative), but also when compared to intra-frame LDA alone. In addition, the use of multi-stream HMMs improves significantly the ASR accuracy, when compared to baseline feature fusion (single-stream HMMs) (Table 2-2), something that becomes more obvious in the next set of experiments.

Table 2-1: Word recognition accuracy for different types of feature selection, namely energy-based, single LDA and two-stage LDA.

Visual feature selection	Video	Depth
Energy-based	36.91%	18.20%
Intra-frame LDA	41.51%	19.28%
Intra and Inter-frame LDA	43.60%	20.72%

Table 2-2: Word recognition accuracy for the single and two-stream HMMs, when using video and depth information.

Fusion Method	Video and Depth
Feature Fusion (single stream HMMs)	41.29%
Decision Fusion (two-stream HMMs)	44.39%

In the second set of experiments, we consider different combinations of the available streams and examine the effects of 3D visual information to ASR when combined with the traditional audio stream. In Figure 2-11 we compare the performance of audio-only ASR to decision fusion based and feature fusion based ASR employing three-stream audio-visual-depth and two-stream audio-depth HMMs. From our results we convey that depth information increases ASR performance, especially for medium and low SNR values. More specifically, depth information increases audio-only ASR by a relative 6% on average (max. 39.7% for SNR=-10dB). In combination with planar video, the relative increase reached 22.1% on average (max. 186.4% for SNR=-10dB). We can also observe, in the case of tri-stream based AVASR, that decision fusion yields significant improvements over the baseline HMMs with feature fusion, specifically 29.3% on average (max. 97.8% for SNR=-10dB).

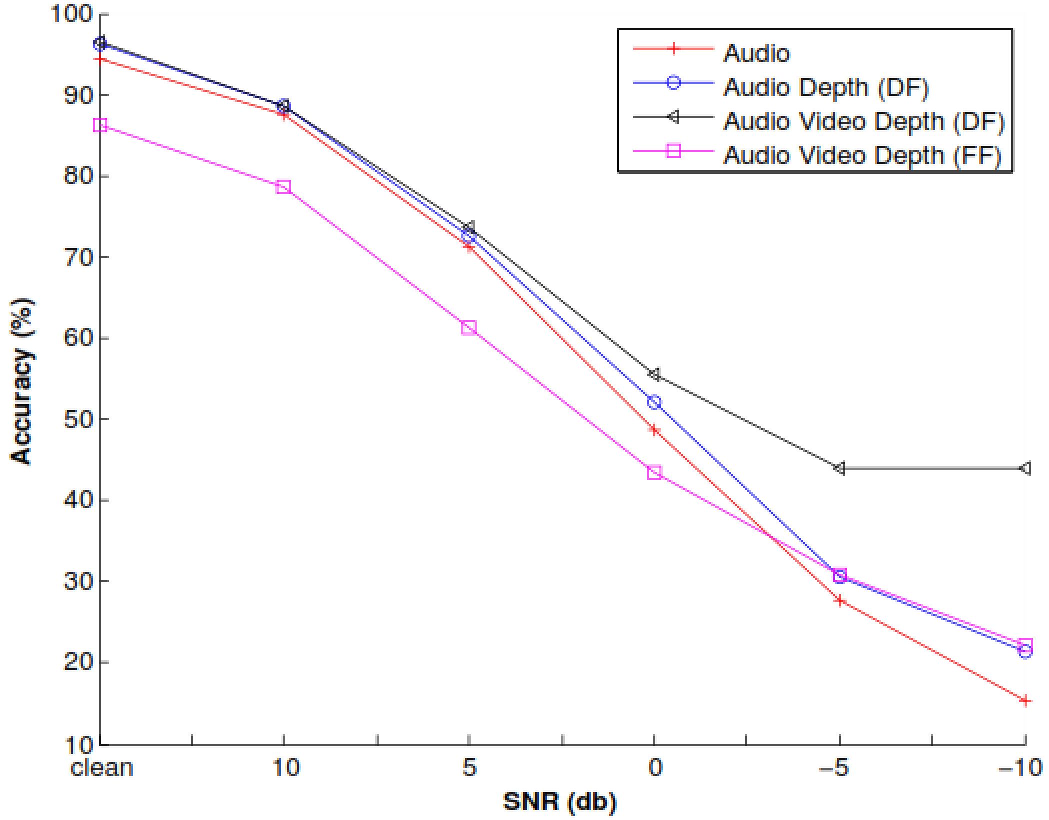


Figure 2-11: Word recognition accuracy results for noisy audio in English using audio-only, feature fusion (audio-video-depth) and decision fusion (audio-depth, audio-video-depth) models.

2.6.1.2 Greek

Additionally, we carried out a set of experiments using the Greek portion of our database. Our experimental paradigm was identical to that used for English, randomly splitting the data from the 6 Greek speakers in a 2/3 and a 1/3 set for training with clean data and testing with noisy data respectively. For these experiments we only used audio babble noise, and intra-frame LDA, while the statistical modeling remained the same. The results attained are shown in Figure 2-12. The highest performance was 99.02% and it was achieved in the lack of noise when all 3 streams were utilized by the system. Lip-reading performance without using audio exhibited a 9.2% improvement with the use of intra-frame LDA. The overall system performance degraded for lower SNR values, but always remained higher than the lip-reading performance. This leads to a significant improvement under very noisy conditions e.g. 45.63% instead of 11.55% word accuracy for our system in comparison to an audio-only recognizer for a Signal to Noise Ratio (SNR) of -10dB.

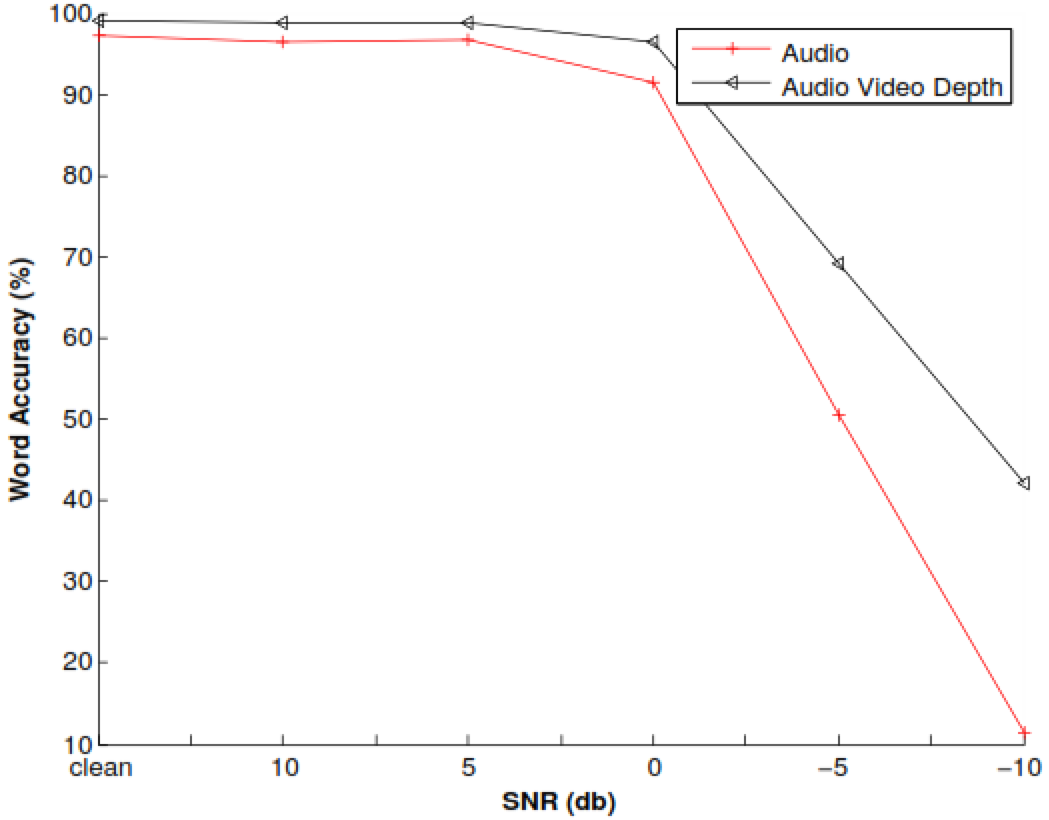


Figure 2-12: Word recognition accuracy results for noisy audio in Greek using audio-only and decision fusion audio-video-depth models.

Comparing with the aforementioned results for English, our system exhibited higher absolute accuracy overall for the Greek language. We believe that this is due to the smaller number of Greek speakers in conjunction with the multi-speaker setup of the experiments.

2.6.2 Noisy Audio and Video

In this section, results carried out on both the English and Greek portions of BAVCD are reported, for: i) audio only, ii) video only, iii) video with depth information and iv) audio with video and depth, at various levels of distortion for each visual noise type. For this set of experiments, audio SNR=0dB. Figure 2-13 to Figure 2-16 summarize the results for English and Figure 2-17 to Figure 2-20 summarizes the results for Greek.

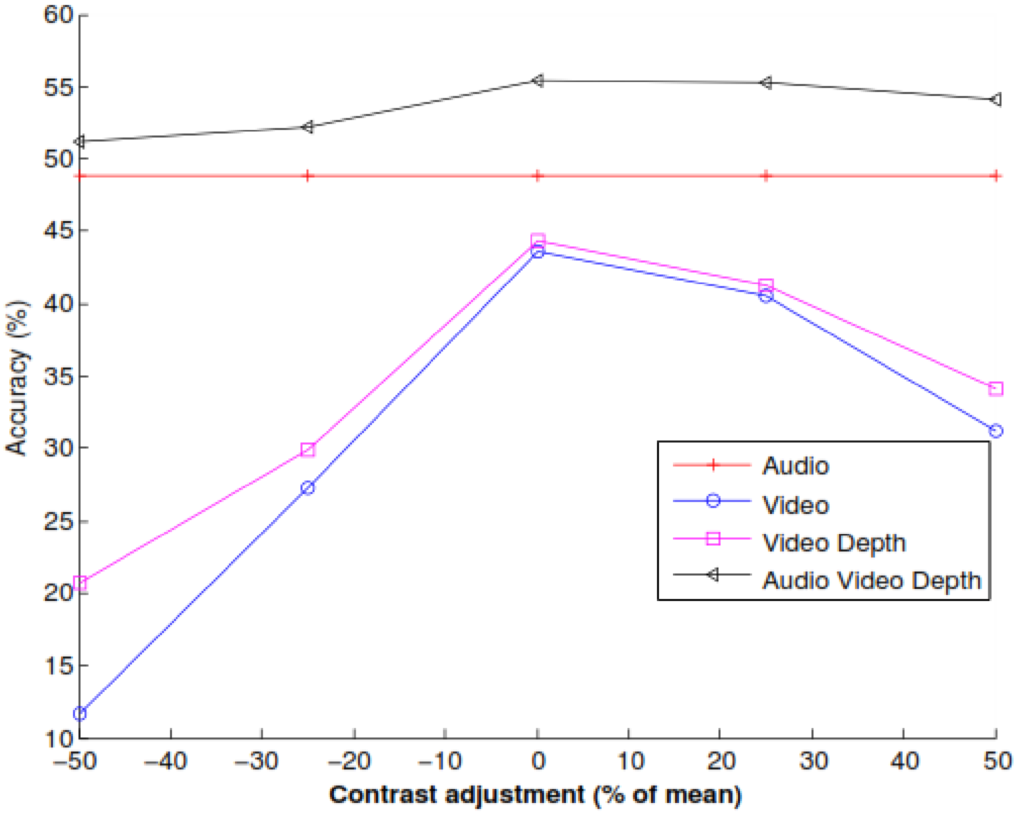


Figure 2-13: Word recognition accuracy results for contrast noise in English.

2.6.2.1 English

The recognition results of our system for a variety of contrast values varying from -50% to +50% of the mean intensity is presented in Figure 2-13. It can be observed that the further the deviation from normal contrast, the lower the recognition performance. As seen in the figure, the use of depth information along with the visual stream helps attain higher recognition rates under all lighting conditions. More specifically the maximum relative improvement was 76.6% for -50% contrast and the average improvement was 13.6%. Furthermore, 3D visual information from the visual and depth streams used with the audio stream attain higher accuracy than the audio-only recognizer. More specifically, the maximum relative improvement was 13.3% for +25% contrast and the average improvement was 9.1%.

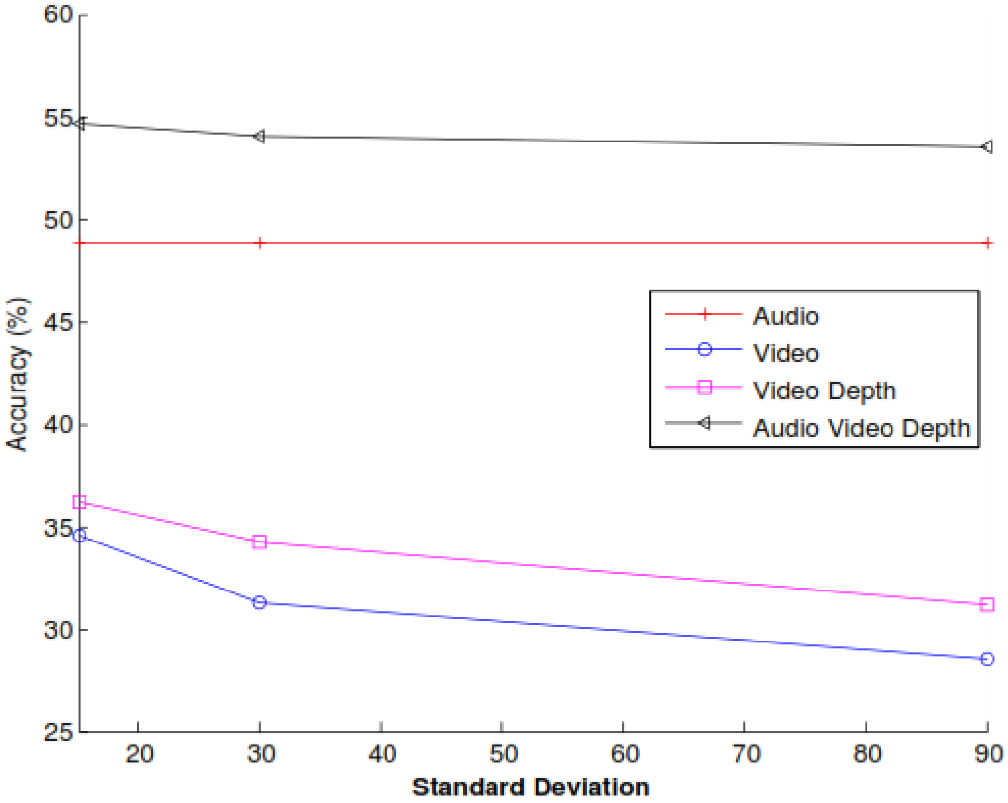


Figure 2-14: Word recognition accuracy results for Gaussian noise in English.

For various levels of Gaussian noise, depth information proves to have a positive impact to the achieved accuracy, mainly for larger degradation levels, as seen in Figure 2-14. In the case of lip-reading, accuracy exhibited a maximum relative increase of 9.6% for std. dev. 30 and on average 7.8%. The combined performance when using audio and 3D visual information increased by 12.1% max. for std. dev. 15 and 10.86% on average. In the same manner, as shown in Figure 2-15 and Figure 2-16, our system exhibits performance gains for both brightness reduction ranging from 75% to 125% of the mean intensity and block noise of 3 different sizes, respectively.

2.6.2.2 Greek

As shown in Figure 2-17 to Figure 2-20, the results for the Greek portion of the database are similar. The main difference between the two languages is the much higher audio-only accuracy for this part of the database at 0dB SNR, which was 91.42% in comparison to 48.78% for the English part. This fact also led to a much higher overall accuracy, when using all three streams. However, the increase in accuracy when combining depth information with noisy video, as well as when utilizing 3-D visual information in combination with audio, remains apparent.

2.7 Conclusions

In this chapter we presented a novel multimodal ASR system that, in addition to the traditional audio and planar video modalities, utilizes facial depth information captured by the Kinect. Feature extraction was employed by means of the 2D DCT and a two-stage LDA feature selection scheme was applied to the visual and depth features in order to boost lip-reading performance. Finally, state synchronous HMMs were used for data fusion and speech recognition. We tested our system under the influence of babble audio noise and 4 types of video degradations and conducted experiments not only in English but also in Greek. Our experimental results demonstrated that the depth modality improves word accuracy in comparison to audio-only and video-only recognition under a decision fusion framework. Furthermore, 3D visual information from both planar video and the depth stream, leads to a significant 22.1% relative increase in accuracy in comparison to a conventional audio-only ASR system. Results when both audio and visual noise are coexistent were consistent for all types of visual noise considered, exhibiting increased robustness with the use of depth information. Finally, word accuracy in Greek appeared higher than English, with reduced dependency from the audio noise level, which we attribute to the smaller number of available speakers.

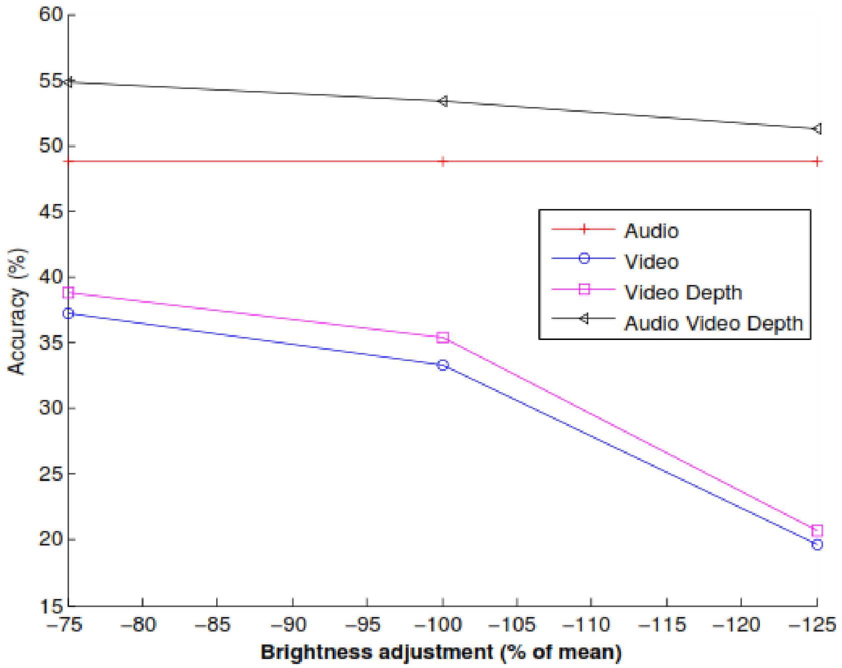


Figure 2-15: Word recognition accuracy results for brightness noise in English.

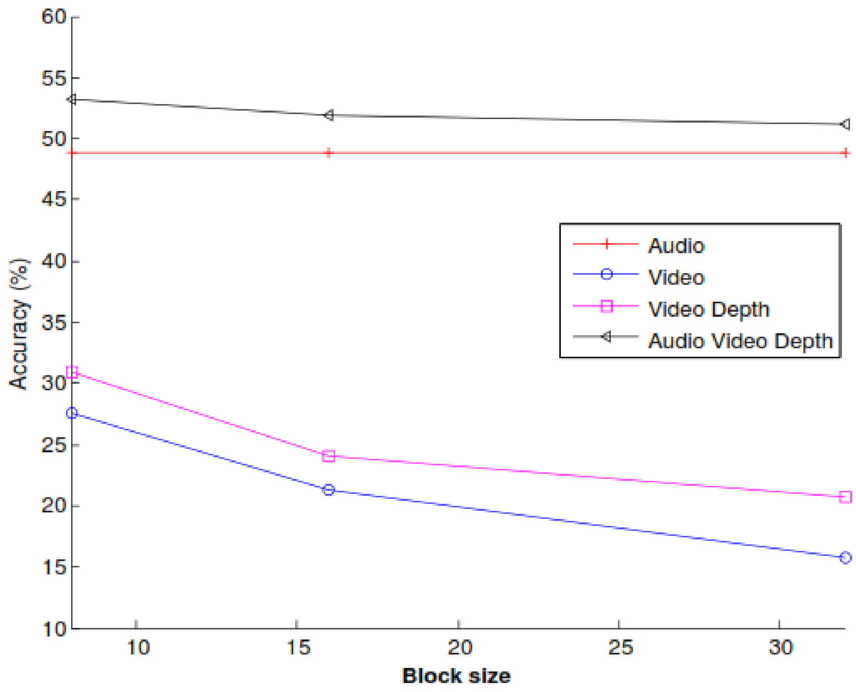


Figure 2-16: Word recognition accuracy results for block noise in English.

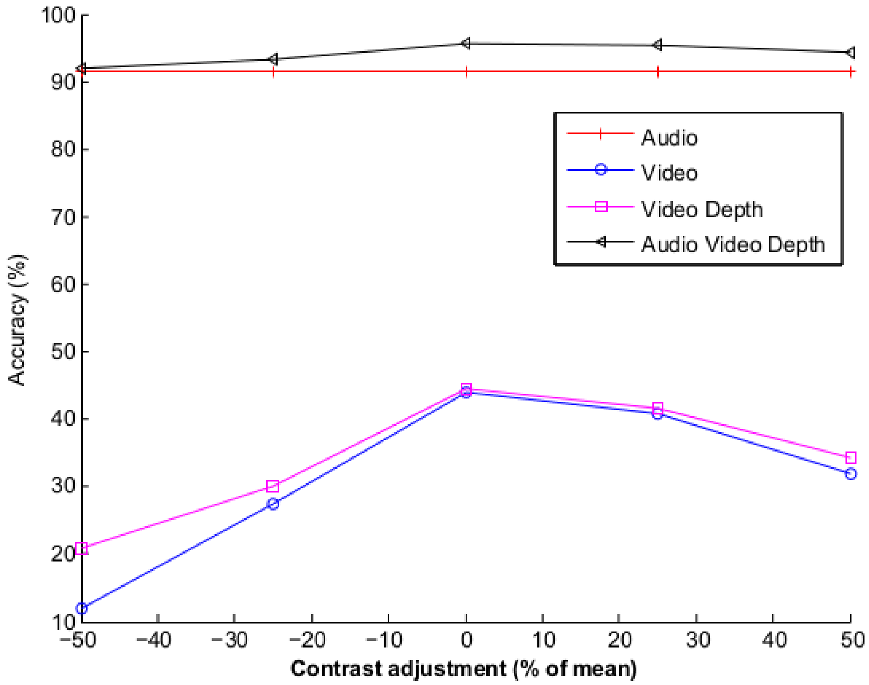


Figure 2-17: Word recognition accuracy results for contrast noise in Greek.

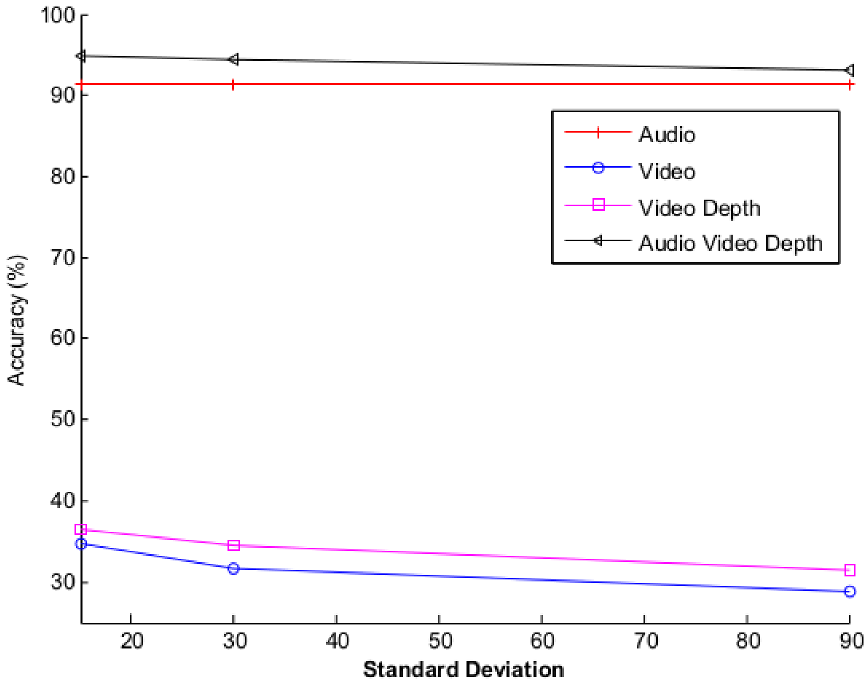


Figure 2-18: Word recognition accuracy results for Gaussian noise in Greek.

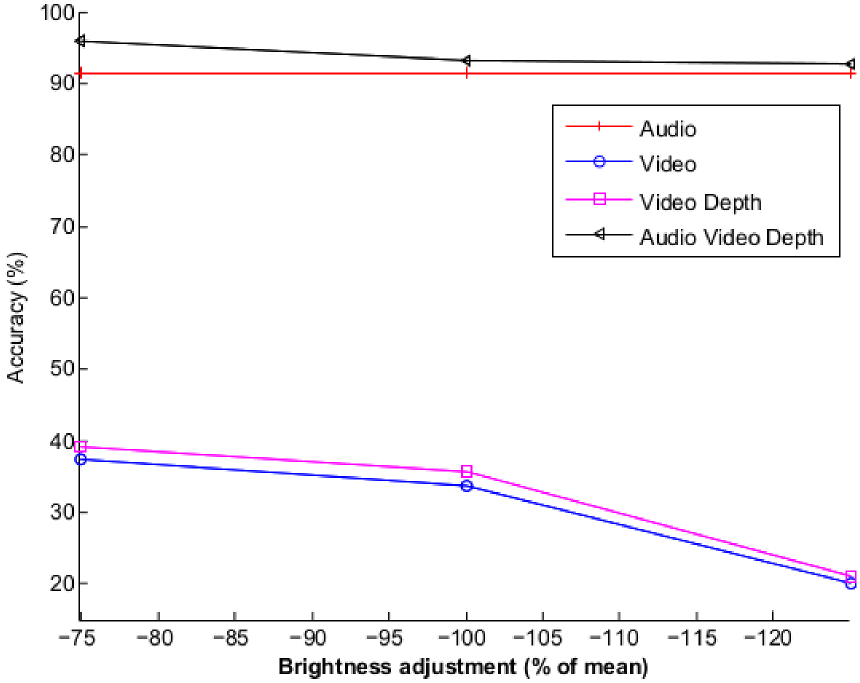


Figure 2-19: Word recognition accuracy results for brightness noise in Greek.

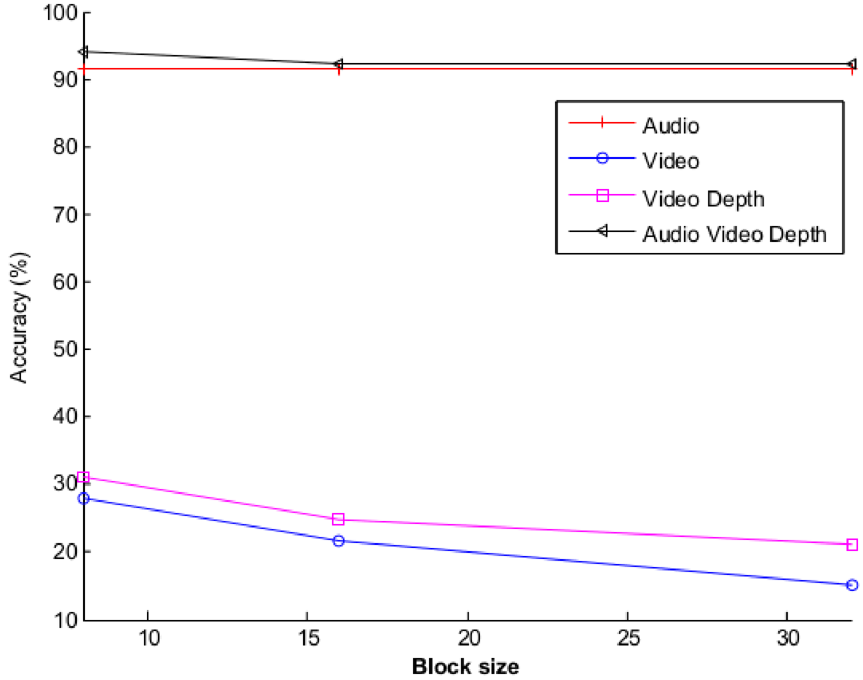


Figure 2-20: Word recognition accuracy results for block noise in Greek.

Chapter 3

Multimodal Identification and Localization

3.1 Introduction and Related Work

As described in Chapter 1, an ambient intelligence environment is a smart space that aids its inhabitants with its embedded technology. The proliferation of ambient intelligent environments has triggered research related to applications, such as monitoring Activities of Daily Living (ADL), fall detection [40] - [44], risk prevention and surveillance [45], [46]. For achieving these goals, activity recognition performed in a natural and unintrusive way is of utmost importance. The most fundamental step towards activity monitoring and ultimately context-awareness is successful multi-person identification and localization. By utilizing the location of the person in a domestic setting, related activities can be derived. Accurate person localization plays an essential role in all the aforementioned applications and has been dealt with using many different approaches. Nevertheless, when used domestically, most current implementations can be considered as invasive.

In this chapter we present the development of a novel system for multimodal, multi-person identification and localization in an ambient intelligence environment. Our unintrusive system uses RFID and 3-D audio-visual information from 2 Kinect sensors deployed at various locations of a simulated apartment to continuously track and identify its occupants, thus enabling activity monitoring. More specifically, we use skeletal tracking conducted on the depth images and sound source localization conducted on the audio signals captured by the Kinect sensors to accurately localize and track multiple people. RFID information is used mainly for identification purposes but also for rough location estimation, enabling mapping the location information from the Kinect sensors to the identification events of the RFID. Our system was evaluated in a real world scenario and attained promising results exhibiting high accuracy, therefore showing the great prospect of using the RFID and Kinect sensors jointly to solve the simultaneous identification and localization problem.

3.1.1 Localization:

Applications that rely on localization such as surveillance and monitoring of 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 [47] and level-set segmentation [48], [49] have been used for tracking, while other approaches based on tracking-by-detection exist [50], [51]. In multi-camera setups, stereo-vision is employed in order to introduce depth perception [52]. In [53], color histograms of the person-shaped blobs are used to disambiguate between people, when they are very close to each other. The system tracks multiple people standing, walking, sitting, entering and leaving in real-time. In [54] two techniques were used to determine the location of a person in 3-D space. These were 1) best-hypothesis heuristic tracking and 2) probabilistic multi-hypothesis tracking to derive the 3-D location of people. The results show similar tracking performance for both approaches. However, the simplistic probabilistic approach produces more false alarms, which may be improved by using a sophisticated probabilistic model.

Solving the problem using only cameras is very challenging for a large space with many people. The reason is that localization requires wide coverage to capture and map the respective locations of many people simultaneously, but identification requires zooming into a person’s face. In surveillance applications, cameras are typically mounted on tall polls and configured such that they could provide maximum coverage.

Nevertheless, video feeds from such settings may not be sufficient to provide accurate information about a person’s face or other biometric features. In addition, the segmentation and tracking problems can be very challenging, thus hindering the system's reliability in a camera-only setup. Furthermore, despite the fact that the use of cameras and computer vision techniques are very promising, extensive use of video cameras in a domestic setting can be considered a violation of privacy [55]. Therefore, our main focus is to achieve the same goal of identifying and localizing multiple people in an assistive environment in a less intrusive manner.

3.1.2 Identification:

RFID (Radio-frequency Identification) systems are frequently being used to track medicine and patients in large hospitals in order to verify the correct medicine reaches the correct patients [56]. RFID sensors have become very popular, as they are cheap, easy to use and provide accurate identification information wirelessly [57]. Although RFID is very effective in identifying objects, it may not be as effective in surveillance applications, since people are required to wear an RFID tag so that the events related to the tag are detected. As a result, such systems may not be able to detect intruders or anyone not wearing a tag. However, it constitutes a viable solution for recognizing activities in a smart environment, since its inhabitants can very easily carry a passive RFID tag with them.

Multimodal person identification has become a significant area of research in recent pervasive assistive applications. Some of these applications use existing biometric identification methods, such as face recognition and speaker identification [55], [58], [59] to identify multiple people in smart environments [60]. Nevertheless, these approaches do not convey the location information of the person.

3.1.3 Simultaneous Identification and Localization:

Locating multiple users simultaneously while identifying each one, is considered to be the first step to create a context-aware application, such as activity and human behavior recognition. RFID technology has also been used to solve the problem of simultaneous identification and localization. Although radio signal propagation suffers from various problems, such as multipath, line of sight path, diffraction or reflection etc. [61] even in an indoor environment [62], several indoor-based localization algorithms have been proposed in the literature, which, according to [63] can be classified into three categories: 1) distance estimation, 2) scene analysis and 3) proximity. Among them, distance estimation algorithms use different range measurement techniques, such as Received Signal Strength, Time of Arrival, Time Difference of Arrival, Received Signal Phase etc. and apply triangulation to estimate the location of the target. On the other hand, the scene analysis approaches first measure fingerprints of an environment and then, try to match the target's range measurements with the appropriate set of fingerprints for estimating the location. Finally, the proximity-based algorithms determine the target's location by mapping it to the location of an antenna that receives the strongest signal.

Overall, RFID technology posses a promising solution to identify and localize multiple objects with attached RFID tags. Existing well-known systems, such as LANDMARC [64] use active RFID tags and exploit the signal strength property to correctly localize an object. Passive RFID tags have also been used in the past to identify and locate multiple objects. In [65], the authors have utilized the percentage of tag counts at different power attenuation levels in order to approximate the distance between a reader and a tagged object. Another, indirect way of deriving the location information of an object is to record the location of the reader as the location of an object. But, the location accuracy and precision of such a system heavily depends on the level of deployment of readers and antennas in the space [66].

However, RFID still lacks sufficient localization accuracy especially for the minimal number of deployed antennas and tags in a domestic environment. Simply using RFID to obtain the location of an object can lead to many false readings, e.g., an RFID antenna may miss a tag depending on the tag's position and the antenna's orientation.

In an attempt to improve accuracy, multimodal person localization has become a significant research area in recent applications. Thus, for a very dynamic environment, information collected from multiple sources, such as video cameras, microphone arrays, sensors etc. are all combined together such that the system can achieve better identification and localization accuracy [73], [74]. Techniques, such as Hidden Markov Models, K-nearest neighbors etc. can be applied to captured audio-visual signals to extract higher-level semantic information, such as identification and location in real time. A system that combines face and audio based identification along with motion detection, person tracking and audio based localization, has been proposed in the literature [75]. Such a system applies state-of-the-art methods to process results from each individual modality and uses particle filtering to fuse both modalities for providing robust identification and localization.

Methods that combine localization using cameras with identification using wearable sensors or accelerometers are also proposed in the literature. Since most of the recent mobile phones contain accelerometers and magnetometers attached to them, mobile phones are considered to be very convenient and fulfill all of the above requirements. In [45] the authors combined an existing CCTV based system with sensors (accelerometers and magnetometers) embedded to a person's mobile phone as a solution. According to this method, the camera captures the location of each person, which is transmitted wirelessly to the mobile phone carried by the respective person. After receiving the location information, the mobile phone resolves the most probable location by matching them with the measurements from its own sensors. The identification process is very easy in this case, as each person is labeled with her/is mobile phone's unique ID.

The deployment of wireless sensor network (WSN) is another common approach nowadays to monitor and localize persons in assistive environments [76], [77]. RFID systems and WSNs can be combined together not only for identifying and localizing objects, but also for real-time monitoring [78]. To identify and localize in open areas, researchers of [46] derived a calibration method for a joint RFID-camera system based on the area of overlap between the field of view (FoV) of a camera and the field of sense (FoS) of RFID sensors.

In our approach, we have utilized the identification capabilities of RFID and combined that with precise 3D tracking from the Kinect to create an accurate identification and localization solution [67] - [70]. The latter is an active sensor, able to accurately measure the position of the person in the 3-D space. Skeletal tracking is carried out using the Kinect sensor's 3D depth images and sound source localization is conducted utilizing microphone arrays of 2 such sensors, to deduce accurate location information. At the same time, the video information is not captured, making this approach less intrusive than using video cameras. RFID is used mainly for discerning between users and also for providing a rough estimate of their location utilizing the RSSI [71]. Our goal is to map the location of multiple people in an ambient intelligence environment at a detailed level that will allow inference of conducted activities (Figure 3-1).

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

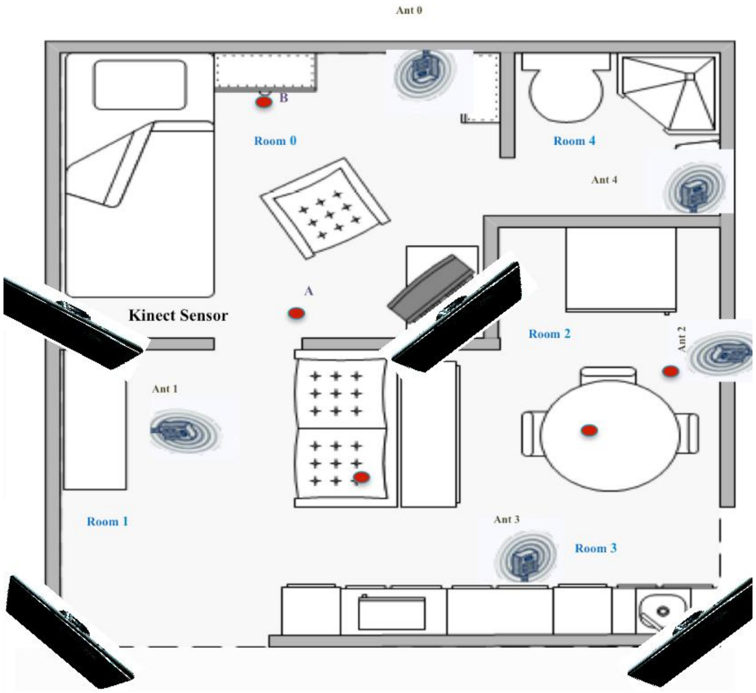


Figure 3-1: Example apartment layout with RFID antennas and Kinect sensors deployment.

3.2 Theory and System Overview

3.2.1 Hardware

The Microsoft Kinect (Figure 2-1) has already been described in chapter 1. In this system we used the Kinect for Windows in conjunction with the Kinect SDK and in addition to its depth sensor, we also utilized its microphone array. The range of the depth sensor is 2.3-20 ft. but it is restricted to 13 ft. by the SDK and 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.

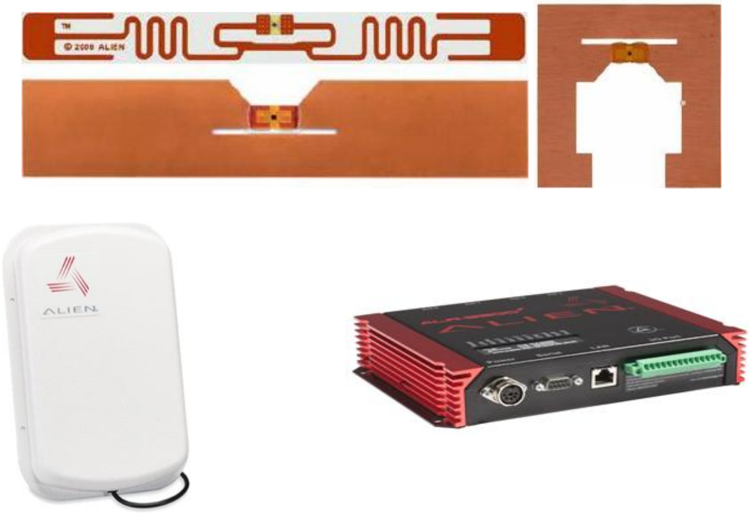


Figure 3-2: The RFID equipment used including different designs of Alien RFID Tags (top) and receiver, antenna (bottom).

An RFID system is comprised of the radio scanner unit / tag reader and the remote transponders / tags. Each tag consists of a microchip transmitter with internal memory and an antenna. The memory architecture may allow the programming or re-programming of the ID or it can be read-only. The 2 RFID technologies in use are either active or passive, defining if the tags have their own power source or not respectively. Passive tags are more common due to their lower cost, smaller size and longer lifetime, despite their smaller range. The most common frequency ranges used by RFID are LF (125-134 KHz) and HF (13,56MHz), although UHF (860MHz) and microwave (2,4GHz and 5,8GHz) tags exist [80]. The RFID system we have used is the commercially available Alien 9900+ developer kit, which includes a reader with two circularly polarized antennas. The tags used in our experiment are EPC Class 1 Generation 2 supported by the 9900 readers. Figure 3-2 shows the RFID equipment used including different tag and antenna designs from Alien. As the antennas are circularly polarized, the tag orientation is not an issue for our experiment. However, for an indoor environment, the antenna read range for the passive RFID tags varies from 20 to 30 ft. Such a read range is sufficient to detect the presence of a person carrying a tag in the simulated rooms of the Heracleia simulated assistive apartment, given the tags are within the FOS of the antennas.

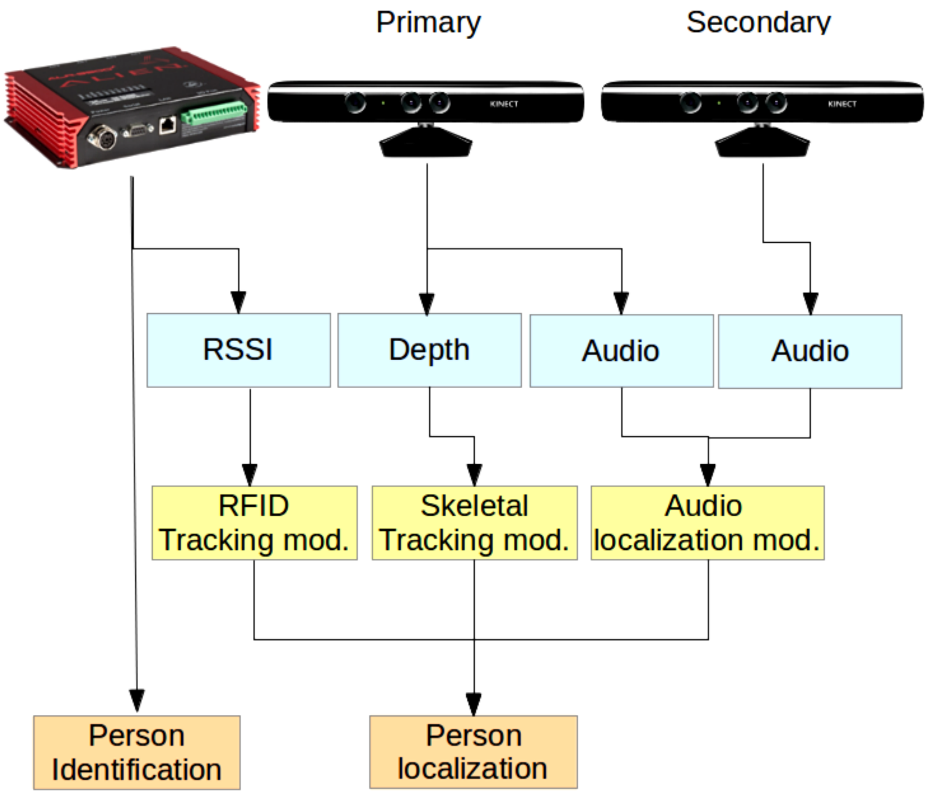


Figure 3-3: System architecture showing the 3 modules and 2 operation modes.

3.2.2 System Architecture

The architecture of our system is modular, comprising of 3 main components as shown in Figure 3-3. Communication between the modules is based around the Joint Architecture for Unmanned Systems (JAUS) [79]. The JAUS architecture is a collection of standards, originally developed by the United States Department of Defense, for unmanned systems. JAUS is designed to govern the way that unmanned systems are designed at the networked component level, as well as the networked agent level. 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 the RFID reader and the 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 [83]. The 3 modules 1) skeletal tracking based localization, 2) audio localization and 3) RFID tracking are described in detail in the following paragraphs.

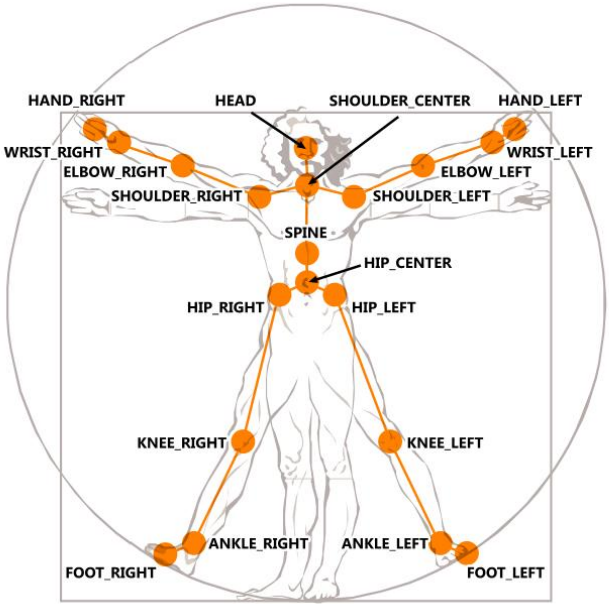


Figure 3-4: The 20 joints tracked by the MS Kinect SDK skeletal tracker [84].

3.2.2.1 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/is 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 $< X_{di}, Y_{di}, Z_{di} >$ of its 20 joints (Figure 3-4) expressed in meters as shown in Table 3-1. 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. The infrastructure for tracking the joints of 2 skeletons and the center of mass of 4 additional people exists, although for the main scope of our system is to monitor an elderly inhabitant of an assistive environment when not supervised, 2 tracked skeletons would suffice. 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 Figure 3-5.

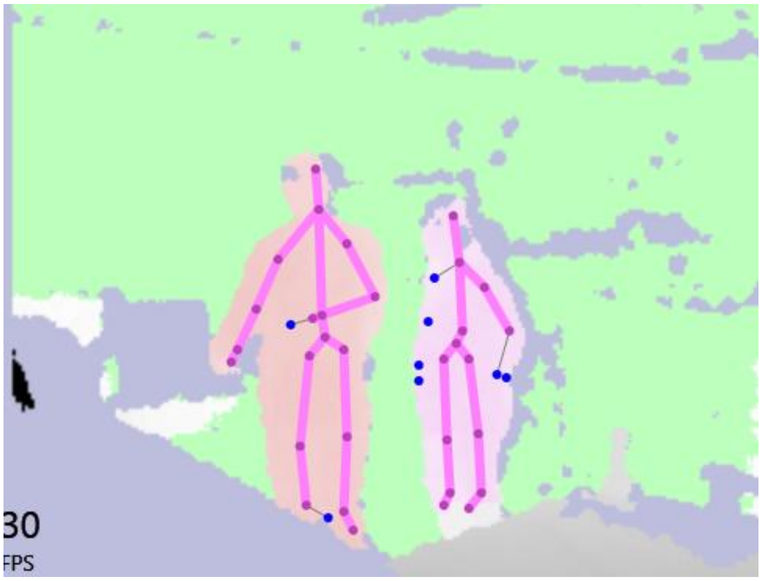


Figure 3-5: An example frame captured featuring skeletal tracking using the Microsoft Kinect SDK.

Table 3-1: Example of data captured by the Kinect skeletal tracker when 2 people are detected in the FOV of the sensor

User	Time	X Cord.	Y Cord.	Z Cord.
1	4/25/2013 1:22:28 PM	-0.1936212	0.1681233	3.099599
1	4/25/2013 1:22:30 PM	-0.08460984	0.08594385	3.164108
1	4/25/2013 1:22:34 PM	-0.4662972	0.07648824	2.894816
2	4/25/2013 1:22:40 PM	-0.5278196	-0.0273742	3.011885
1	4/25/2013 1:22:45 PM	-0.4450822	-0.1373514	2.829979
2	4/25/2013 1:22:50 PM	-0.456997	0.07926513	2.96067
1	4/25/2013 1:22:53 PM	-0.4790949	0.08204032	2.961271

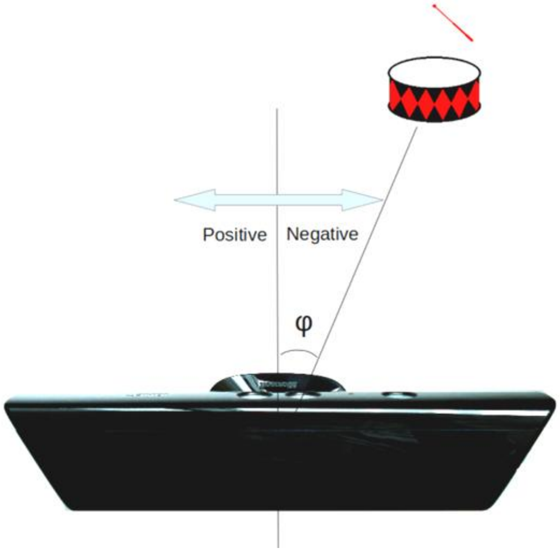


Figure 3-6 Kinect sound source localization and detected sound source angle sign convention.

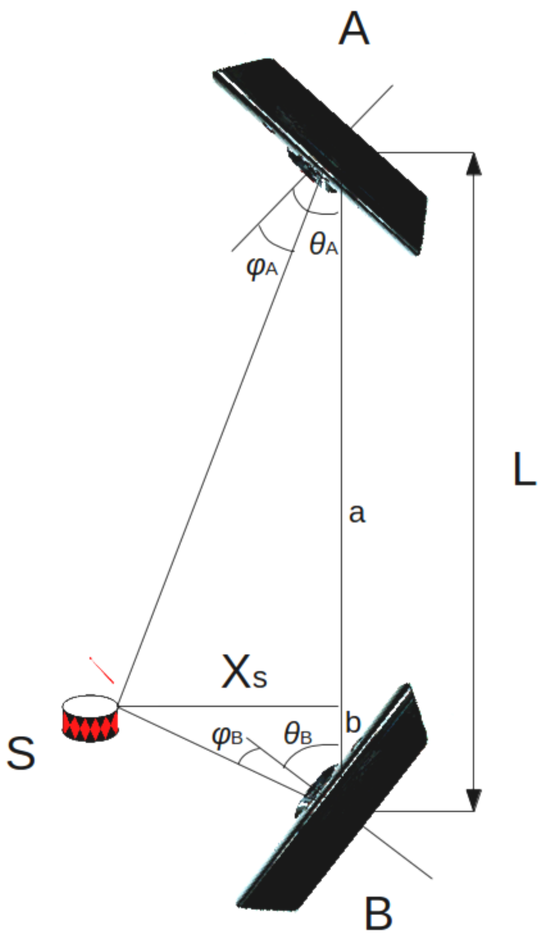


Figure 3-7: Configuration of the 2 Kinect devices for audio localization of an audio source on a 2-D plane.

3.2.2.2 Audio Localization Module

The microphone array of the Kinect is comprised of 4 super-cardioid 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 beam-forming 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 (Figure 3-6). 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.

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 (Figure 3-7). 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 $\theta_{\mathrm{A}}, \theta_{\mathrm{B}}$ 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 $\phi_{\mathrm{A}}, \phi_{\mathrm{B}} \in (-50,50)$. 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 (Figure 3-6). 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:

$$\tan(\theta_{\mathrm{A}} - \phi_{\mathrm{A}}) = \frac{X_s}{\mathrm{a}}$$

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

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

$$X_s = \frac{\tan(\theta_{\mathrm{A}} - \phi_{\mathrm{A}}) \cdot \tan(\theta_B + \phi_B) \cdot L}{\tan(\theta_{\mathrm{A}} - \phi_{\mathrm{A}}) + \tan(\theta_B + \phi_B)}$$

$$\mathrm{a} = \frac{X_s}{\tan(\theta_{\mathrm{A}} - \phi_{\mathrm{A}})}$$

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

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.

3.2.2.3 RFID Based Localization Module

The RFID system that we used was comprised of two antennas and a tag reader. Its main role was to identify the person in its field of sense (FOS), but also to provide a rough estimate of her/his location using the received signal strength indicator (RSSI) from each antenna (Table 3-2). The mapping between the RSSI values and the actual position of the tag is accomplished through a calibration process that accounts for both the directionality of the antennas and the specific layout of the room. Multiple people are identified using their unique RFID tag and tracked as long as they remain in the FOS of the system. Skeletal tracking alone may not be able to discern between different people since a new tracking id is issued each time a person is lost from the FOV of the Kinect and then re-enters. Therefore, we improved our system's accuracy by matching the new RFID tag with the new tracking id as soon as an individual enters the room. This technique allows identification of each individual detected by the skeletal tracker. In the case where an unmatched tag id or skeletal id appears e.g. if a person was not detected upon entrance by either sensor, they are matched when they both appear in the same sector. Finally, when no skeleton is detected in the FOV, but a tag is still being detected, audio localization is utilized in order to increase accuracy (e.g. when only one antenna reads the tag).

Table 3-2: Example of RSSI data as captured by both antennas of the system when 2 tags are detected in the FOS.

ID	Time	Antenna	RSSI
E2009037890401091080A8BA	4/25/2012 1:22:52 AM	1	4792.4
E2009037890401090900BD64	4/25/2012 1:22:52 AM	2	1351.6
E2009037890401090900BD64	4/25/2012 1:22:53 AM	1	1021.2
E2009037890401091080A8BA	4/25/2012 1:22:54 AM	1	4920
E2009037890401090900BD64	4/25/2012 1:22:54 AM	2	1299.1

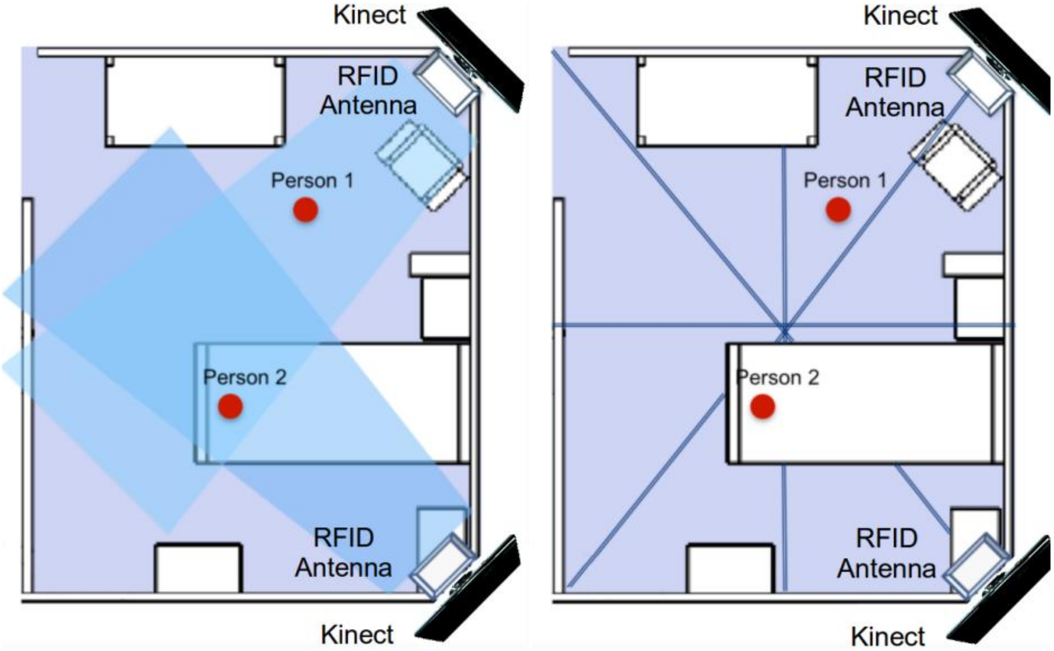


Figure 3-8: Layout of our simulated apartment for person identification and localization combining 2 RFID antennas and 2 Kinect devices (left) and division of the apartment into 8 sectors for localization using RFID (right).

Localization is based on a training phase during which pre-specified position signatures (RSSI in our case) are used. More specifically, we divide the entire room into multiple sectors, as shown in Figure 3-8. Next, we collect the RSSI signatures of the detected tags in these different sectors using the antennas. In the training phase, we label the signatures with their corresponding sector number and a model is fitted to our data in order to describe the relationship between the location and the observed values as well as to predict the location for new values. This way we build a classifier that classifies any RSSI measurement from an antenna into one of these different sectors. The idea is that given that an RFID tag is detected, the system first narrows down its location to one of the sectors. Next, given the measurement from the Kinect sensor for any particular person, if the measured location falls within that specific sector, then we map that particular person to the location described by the Kinect sensor. As aforementioned, in both approaches we use the sound from the microphone array as another modality besides skeletal tracking to resolve ambiguities in mapping.

3.3 System Operation

The main function of our system is person localization utilizing information from all three modules. 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, as described in the previous section.

In order to determine the final estimated location of the person we consider the available localization information from all three modules hierarchically, according to our experimental results presented in the next section. So, in the case where one of the modules does not return any coordinates, then the other module's coordinates are considered. The order in which we determine the location of each person is: 1) Skeletal tracker, 2) RFID, 3) Sound source localization. If skeletal tracking information becomes unavailable (e.g. if the person is outside the FOV of the depth sensor), then the system relies on RFID. Similarly, if both skeletal tracking and RFID information are unavailable (e.g. tag undetected by 1 antenna), then sound source localization is used. In addition, we experimented by calculating the average location for each person. More specifically, when a location estimate is available from both the RFID and the skeletal tracker, 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. 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.

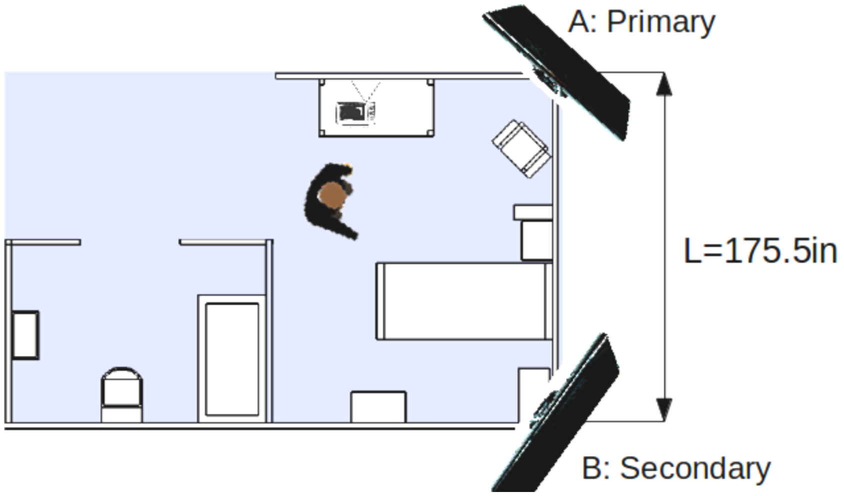


Figure 3-9: Simulated apartment layout and placement of the Kinect devices.

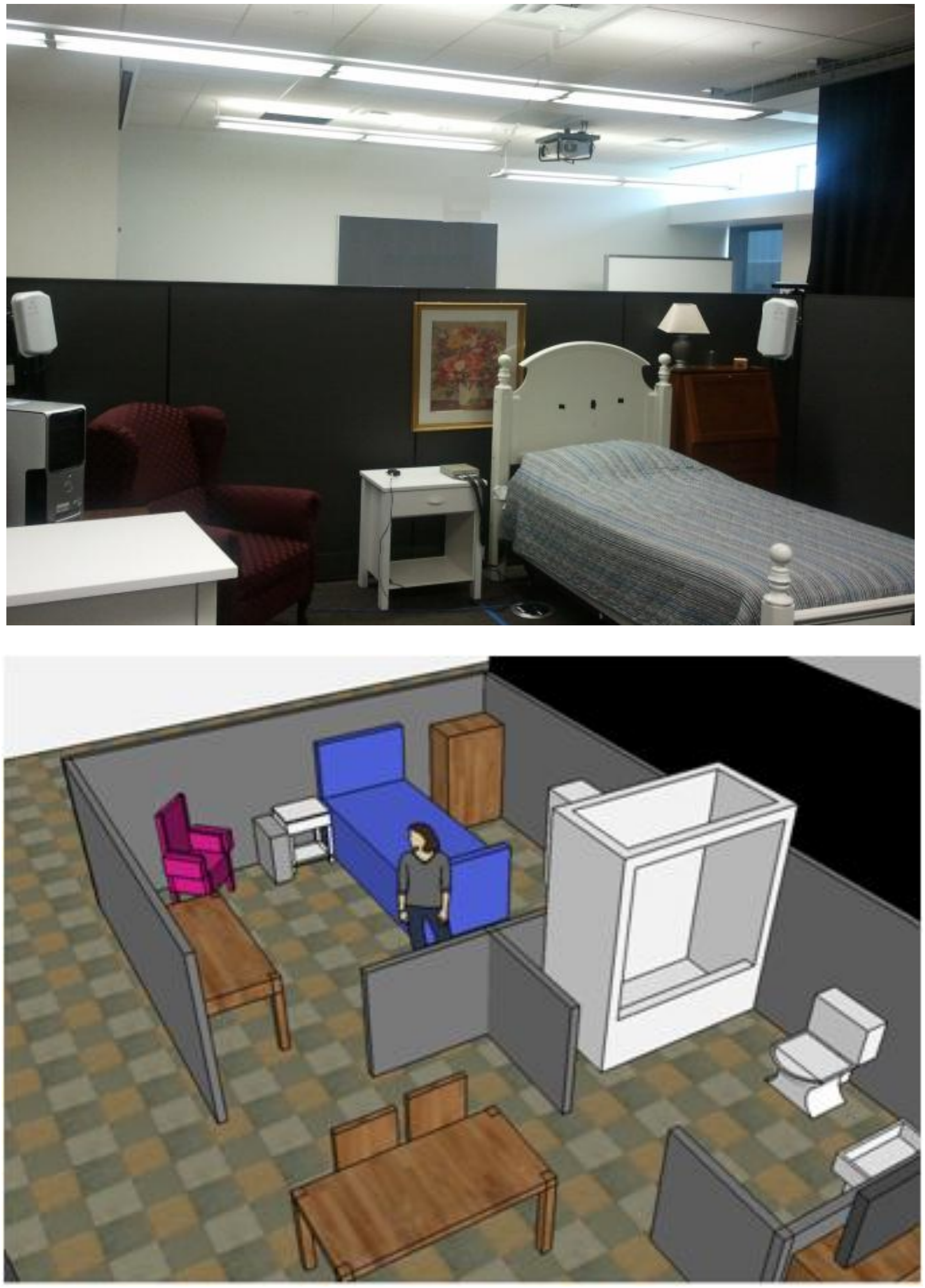


Figure 3-10: An aspect of our simulated apartment (top) and a 3-D reconstruction of our simulated apartment (bottom).

3.4 Experimental Setup

An extensive set of evaluation experiments were conducted in order to fine-tune the parameters of the setup at our simulated apartment . As mentioned earlier, two Kinect devices and two RFID antennas 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 sensors' axes pointed at 45 degrees towards the interior of the apartment, maximizing both the FOS, FOV and microphone coverage (Figure 3-9 and Figure 3-10).

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.

Table 3-3: Experimental results for the identification and localization tasks for all three modules.

Task/Source	**4 people**	**2 people**	**1 person**
Identification/RFID	92.5%	97.1%	100%
Localization/Kinect	90.3%	93.8%	98%
Localization/RFID	82.1%	87.2%	85.4%
Localization/Sound	75.9% (1 speaker)		

In our experimental setup, we have deployed two antennas at the two corners of the bedroom, as shown in Figure 3-8. We have simulated an experiment for identifying and localizing up to 4 people, limited only to part of the apartment, although the system can be extended to more rooms by adding more Kinect sensors in the apartment. During the experiment, each person wears an RFID tag around her/is neck.

We conducted extensive experiments in our realistic domestic setup. Four individuals participated in our experiments, with one, two or four occupying the apartment simultaneously. Subjects were asked to move in the apartment in 10 sessions and perform 4 activities, namely walk and sit in a chair, at a desk or on a bed. The total number of instances used for the classification was 400 per experiment. In Table 3-3: Experimental results for the identification and localization tasks for all three modules. we report results for both the identification and localization tasks after 10-fold cross validation. For both tasks, accuracy degraded for more occupants, due to the people interacting and the resulting occlusions. Identification accuracy using RFID was at very high levels, considering single antenna misdetections. Localization accuracy denotes the percentage of correctly estimated locations for all individuals present in the room and also accounts for misidentifications and mismatches between the detected tag sector and skeletal id location. The accuracy attained using the Kinect was over 90%, and constituted the most accurate source for person location information. The accuracy achieved using RFID was over 80% and 75.9% using sound (only 1 speaker).

3.5 Conclusions

In this chapter we presented a novel system capable of accurate and robust person localization and identification. Our system combines the tracking capabilities of the Kinect sensor with identification information from existing RFID technology. 3 types of data were used to solve the localization problem, namely the RSSI, 3D depth and audio information. Accurate position estimation for each person was carried out using the depth sensor and microphone array of the Kinect as inputs, by means of skeletal tracking and sound source localization respectively. The system was deployed in a simulated apartment and during the experiments conducted, it achieved high localization and identification accuracy, proving its effectiveness for the 4-person localization scenario.

Chapter 4

Assistive Robot Navigation and Placement

4.1 Introduction and Related Work

Person localization, identification and activity monitoring in general are very important in an ambient intelligence environment. Nevertheless, a dynamic environment that changes over time, requires the adaptation of the sensors' configuration. Furthermore such a system must compensate for sensor failure in an online fashion, where the remaining operational sensors compensate for the ones that failed. Finally, this system must achieve maximum coverage of the monitored space for increased efficiency and cost containment. The field of robotics in particular has shown great potential in addressing these challenges.

In this chapter, we present the development of a novel system that optimally positions a number of robots in an assistive environment and can be used in a domestic environment inhabited by an elderly, disabled or visually impaired person, for guidance and event recognition [81]. Placement of the robots is done by means of the Extended Max Sum Decentralized Coordination algorithm. The sensor readings are taken into account during the optimal placement process, in addition to the environment layout, in order to maximize the system's effectiveness. Our framework was built on top of our previous work, the prototype assistive-guide robot eyeDog [82], shown in Figure 4-1, which was initially developed to provide the visually impaired with autonomous vision-based navigation and laser-based obstacle avoidance capabilities. Therefore, we also describe its design and development. This kind of assistive-guide robot has several advantages, such as robust performance and reduced cost and maintenance. The main components of our system are the Create robotic platform (from iRobot), a net-book, an on-board USB webcam and a LIDAR unit. The camera is used as the primary sensor for the navigation task; the frames captured by the camera are processed in order to robustly estimate the position of the vanishing point associated to the road/corridor where the eyeDog needs to move. The controller then steers the robot until the vanishing point and the image center coincide. This condition guarantees the robot to move parallel to the direction of the road/corridor. While moving, the robot uses the LIDAR for obstacle avoidance. The novelty of our implementation is the use of RANSAC with adaptive thresholds to estimate the vanishing point of the visual scene as well as the fusion with the laser data in order to navigate the robot. The feasibility and effectiveness of the approach is demonstrated by successfully guiding the user down the center of a path, such as a hallway or sidewalk, while navigating around obstacles and walls.

The development of an effective aid, especially for the disabled or visually impaired is a demanding task which requires multiple challenges to be resolved [85]. The nature of such a condition requires that both location and situational awareness be provided, while conveying this information to the user effectively and in real-time. Perhaps the most identifiable and useful aid for the visually impaired is a specially-trained guide dog. Guide dogs are trained to lead their owner through the environment, remaining on a given path, while avoid obstacles and hazards, increasing her/is safety. While the benefits of guide dogs are undeniable, many factors make them unsuitable or impractical in several situations. Guide dogs require extensive training before they can be matched with a user. Once a suitable dog has been identified and trained, the person must also be trained to operate as a single unit with the dog. Perhaps the greatest limiting factor in guide-dog use is their cost. The average cost of a guide dog in the United States, including the necessary training for both the dog and the user, is $42,000 [86]. Other limitations include allergies, location and availability of training schools, time required for training etc. which make guide dogs impractical in many cases. The use of robotics is a promising alternative to guide dogs. A properly designed robot-navigation aid could be both performance and cost effective. Recent advances in sensing technology, particularly in computer vision, allow a robot to identify paths and objects, thus enabling effective location and hazard recognition.

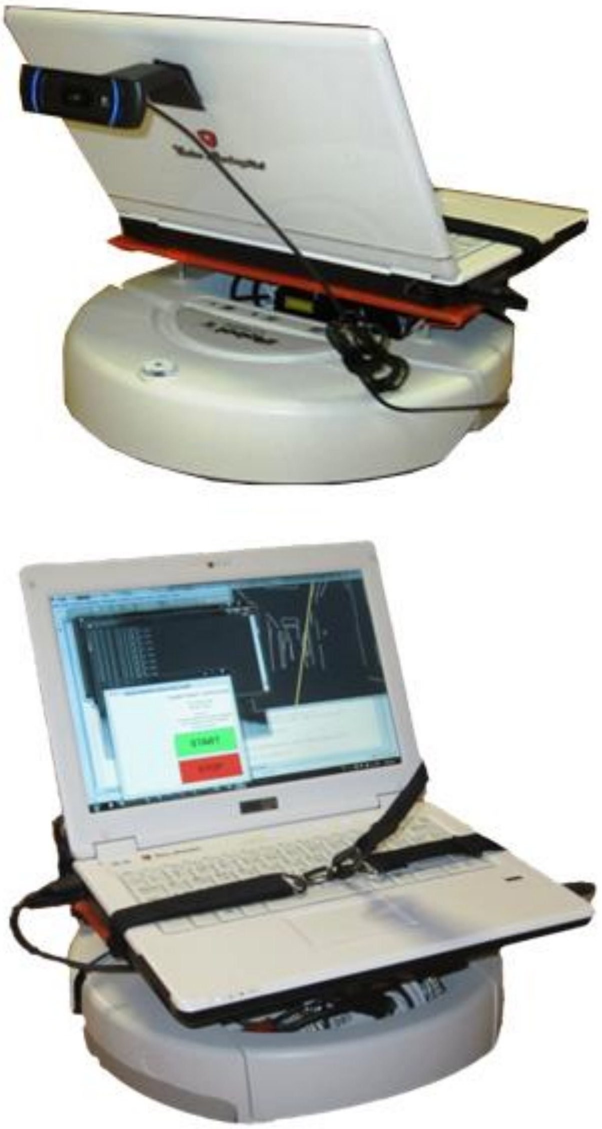


Figure 4-1: eyeDog, the prototype assistive-guide robot.

The problem of guiding a robot along a certain path or corridor belongs to the broader range of lane-detection problems. These types of problems have been attacked using various approaches during the last years and even competitions like the DARPA Urban Challenge [87] or the Mini Grand Challenge [88]. There have been different approaches for lane detection, including particle filters [89] and homography [90]. One of the most popular approaches consist of using RANSAC for robust clustering of multiple lines in the image space in order to determine the vanishing point in an image [91] - [94].

However, research efforts have focused on car traffic and not assistive robots. To our knowledge there has been only one previous attempt of an assistive guide robot for the visually impaired, which focused on indoor navigation. This was developed in the Utah State University by Kulyukin et al. [95]- [97] and was built on a Pioneer robotic platform, utilizing a SICK LIDAR for obstacle avoidance, a webcam, and RFID in order to locate a specific product in a store and then navigated to its position. The navigation was accomplished using a colored tape on the floor of the aisles and RFID tags on the intersections. This system requires a number of modifications in an existing store in order to be deployed, such as RFID tags on all product packages and placing colored tape on all aisles. In contrast, our approach [82] uses Computer-Vision techniques to navigate a specific path without requiring any changes to the environment. Furthermore, we expanded our previous work, aiming at utilizing a multi-robot configuration framework that is able not only to navigate throughout an apartment but also monitor and recognize certain events such as loud noises and movement. The existence of more than one robots can ensure more effective monitoring capabilities and responsiveness in case of need by the user. In such a case, the problem of positioning the robots in the apartment is of utmost importance, since a fully automatic and efficient algorithm must be used so as to ensure effective and optimal coverage of the monitored space by the robots' sensors, while taking into account both the layout of the space and the characteristics of the sensors used. We thus utilized a tool for sensor placement and system monitoring that allows for designing the environment layout, defining crucial areas and updating the position of the robots dynamically. We followed a similar approach, presented in [104] where we used cameras and a decentralized framework in the form of a multi-agent system. In our approach we have incorporated more types of sensors to our robot which is also able to move in the apartment [81].

4.2 Theory and Robot Architecture

The eyeDog is comprised of the following main components: the iRobot platform, a notebook computer, a Logitech USB webcam and a Hokuyo LIDAR unit shown in Figure 4-2. The camera is the primary sensor for the navigation task and is used to estimate the vanishing point from the captured video. This video sequence is processed using OpenCV 2.1 [103] which extracts prominent lines in the image. Then RANSAC is used to determine the most probable vanishing point. After the vanishing point has been determined, the deviation from the principal point of the camera is calculated and the robot steers accordingly in order to move parallel to the direction of the road. While moving, the robot uses the LIDAR scanning system for obstacle detection and avoidance.

In addition to the navigation task, we provide a modular software development platform to facilitate the evaluation of a variety of potential sensing and control approaches. This “plug and play” philosophy is loosely based around the Joint Architecture for Unmanned Systems (JAUS)[79] (see Chapter 3), originally developed by the U.S. Department of Defense, to govern the way that unmanned systems are designed. While a fully compliant JAUS implementation is out of the scope of this project, the software architecture utilized in the robotic guide dog follows many of the guidelines established by JAUS (distribution of software modules, UDP communication between modules, etc). This increases the level of interoperability at the component level, allowing a new software module to be quickly and easily integrated in the system without changes to other components.

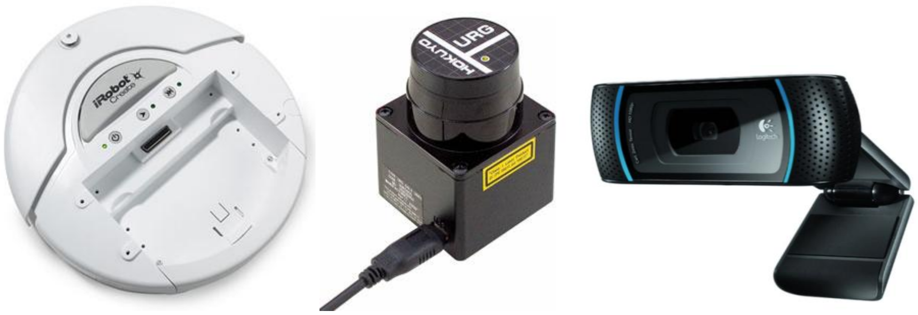


Figure 4-2: The hardware components of the eyeDog: the iRobot platform (left), Hokuyo LIDAR (middle) and Logitech webcam (right).

4.2.1 Control Software Architecture

As mentioned previously, the architecture employed by the system is distributed at the component level. Individual software modules were created for specific tasks, such as polling a particular sensor or sending actuator commands to the iRobot platform. These background processes communicate over a network protocol using UDP packets in a specific format. The modules are able to run on the same machine or on their own dedicated hardware simply by specifying the IP addresses and port numbers accordingly. This networked approach makes it possible to distribute computational load across several computers for increased performance and expandability. The software architecture employed by the eyeDog is shown in the diagram in Figure 4-3 inside the “control” box.

Each software module is assigned a unique IP address and port number pair, allowing them to be individually addressed across multiple hardware units.

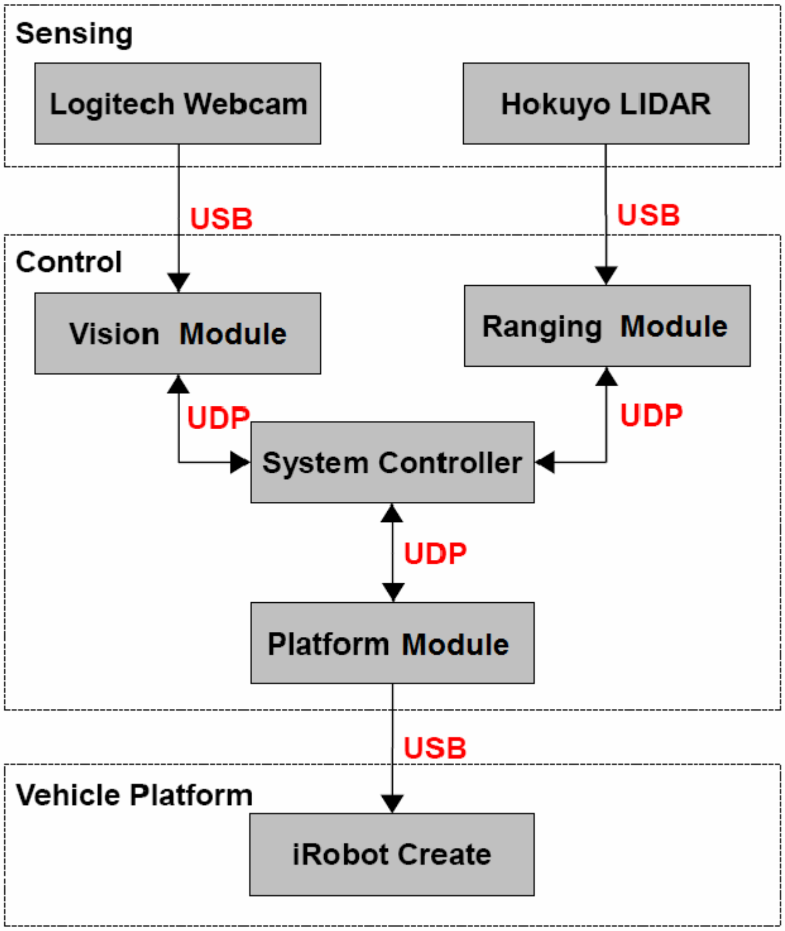


Figure 4-3: Architecture of the prototype eyeDog robot, displaying the communication between the sensing hardware, control software modules and vehicle platform.

The vision and ranging modules monitor the camera and the LIDAR, respectively, compute error signals, and send perceived values to the system controller. The system controller implements the robot's PID controller, provides error reference signals to the sensing modules, and visualizes the incoming data streams. Every aspect of the vehicle control law (such as the period, PID gains, and filtering parameters) is configurable from the system controller. Filtering of the incoming data streams is performed independently, with both a moving-average and moving-median filter implemented. These filters, running in parallel, can be modified during execution to aid in testing and performance evaluation.

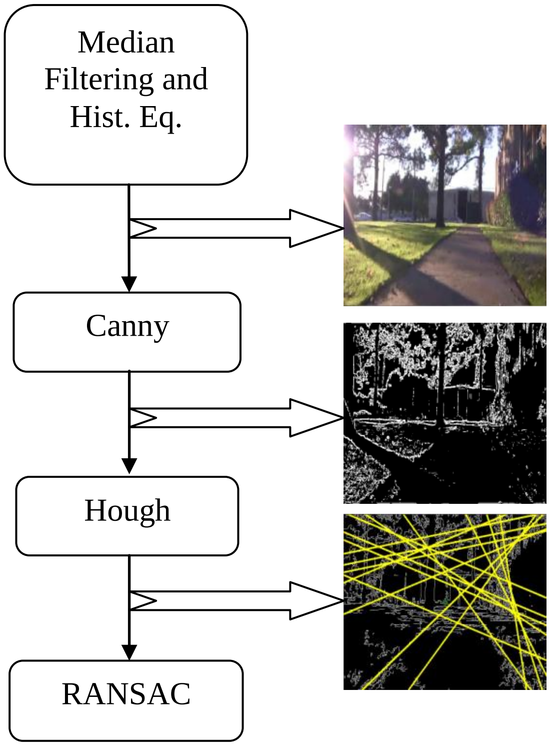


Figure 4-4: Vision module pipeline and example.

4.2.2 Vision Module

The vision module is responsible for estimating the pixel position of the vanishing point at each time instant. Initially, image processing techniques are used based on the Canny edge-detector in conjunction with Hough transform to extract the prominent lines in each image. Since the effectiveness of this feature extraction step has been proven, we decided to also use it in our assistive robot application. The main vision pipeline is illustrated in Figure 4-4. After the current frame is obtained, the Canny Edge Detection algorithm is used to extract all the image edges. The Hough transform is then adopted to extract only the most prominent image lines. Finally, RANSAC is adopted to detect the position of the vanishing point while simultaneously extracting those image lines parallel to the road.

4.2.2.1 Canny Edge Detection

The first step of image processing is the detection of edges in each frame of the captured video. In order to improve the performance of our edge detector, each image has to be preprocessed. Initially the image is converted to grayscale, since Canny operates on this kind of images. Secondly, histogram equalization is performed on the image in order to increase contrast and therefore emphasize edges [98]. Histogram equalization is an image enhancement technique that operates on the spatial domain. It modifies the distribution of the pixels to become more evenly spread out over the available pixel range. Since a histogram of a grayscale image displays the distribution of the pixel intensity values, histogram equalization attempts to reshape the probability distribution function (PDF) into a uniform function (Figure 4-5). Therefore, although a dark image will have mostly low intensity pixels and a bright image only high intensity pixels thus lowering the contrast, an image with a uniform PDF will have pixel values at all valid intensities [99].

Before proceeding to the actual edge detection, a median filter operates on the image. This is a smoothing filter that causes blurring but at the same time preserves larger edges. This aims at adjusting the size of edges that we want to preserve by choosing a corresponding filter size. In our application the median filtering uses a window of 5x5.

After preprocessing, the Canny edge detector is used to extract the edges of the image. The operation of the Canny edge detector in general can be summarized as follows. Normally, a Gaussian blurring filter is first convolved with the image in order to discard noise and unimportant edges, although this is not performed in the OpenCV implementation of Canny. Then, a Sobel operator is applied in order to calculate the gradient norms in the x and y axes. This is followed by hysteresis thresholding which uses 2 thresholds in order to determine whether a pixel is an actual edge pixel, discarding small edges and noise. The higher threshold for the canny edge detector is set adaptively using Kerry Wong's [99] formula high_thresh=1.66*mean, where mean is the mean value of the pixels' intensity after the image has gone through histogram equalization. The lower threshold has been defined using the empirical rule low_thresh=0.4*high_thresh. Finally non-maxima suppression is applied on the image in order to reduce the edges thickness to 1 pixel to clearly define the contours in the image.

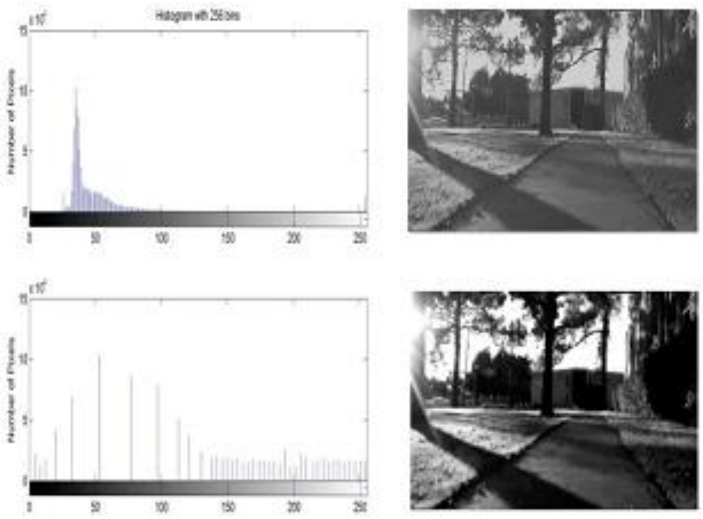


Figure 4-5: Histograms. The original histogram (top) compared to the equalized histogram (bottom). The range of the pixel intensity values becomes broader.

4.2.2.2 Line Localization

Since our ultimate goal is to determine the vanishing point in the image, we first need to detect the perspective lines that intersect at the vanishing point. Therefore after acquiring the edges of the given image, we use the Hough transform [99] of the road to detect the most prominent lines. Hough transform is a voting algorithm that can be used to determine whether there are enough pixels to form a particular shape in the image, in our case a particular line. In order to accomplish that, each line has to be expressed in polar coordinates (ρ θ), so that a generic point (x,y) (pixels) belonging to a line will satisfy the following equation:

$$x\cos(\theta) + y\sin(\theta) = \rho$$

Where ρ represents the distance from the origin to the line along a vector perpendicular to the line and *θ* is the angle between the x-axis and the vector perpendicular to the line (Figure 4-6(top)).

By using this transform, a line can be represented by a single point in the polar-coordinate parameter space. Similarly, since infinite lines pass through any given pixel in the original image, the representation of a pixel in the parameter space is a unique sinusoidal curve (representing all the lines that can pass through that pixel). The point of intersection between multiple sinusoidal curves in the parameter space represents the line passing through all the pixels. Therefore the more the intersection points, the more pixels a line passes through (Figure 4-6(bottom)). The parameter space is then divided into bins in the ρ and θ space. The total number of intersections in each bin is then saved into the accumulator, and the highest voted lines are returned. In addition, only a maximum number of 30 lines with the higher accumulator values that result from the Hough transform are considered in order to save computation time.

After acquiring the parameters of the lines detected in the image, the following filtering procedure is conducted in order to discard lines that cannot be considered for the vanishing point estimation. Since the camera is mounted at a certain height, we assume that the edges of the path cannot appear as horizontal lines. In addition we assume that the robot remains in the path, and therefore there can be no vertical lines coming from the edges of the path. Thus we filter lines that deviate by 5 and 10 degrees from the vertical and horizontal axes, respectively.

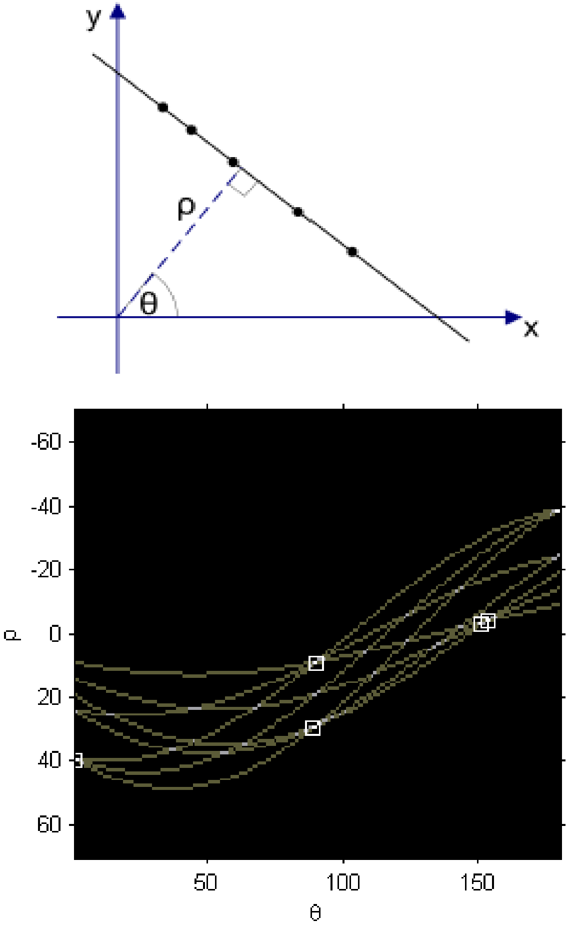


Figure 4-6: Polar-coordinates (ρ,θ) representation of a straight line (top). Each line has a unique representation (ρ,θ). The Hough space of an image (bottom), indicating the points with the highest number of intersections. Many points are around 90 degrees, i.e. the image has many horizontal lines.

4.2.2.3 RANSAC

The result from the line localization step consists of a set of lines in the image at a specific slope and distance to the image center. We will assume hereafter that most of the lines in the image will pass near the vanishing point. Under this assumption, we can use RANSAC (RANdom SAmple and Consensus) [94] in order to randomly sample the above set of lines and find an estimate of the point of intersection (i.e., the vanishing point) which has the highest consensus. RANSAC is a non-deterministic iterative method for estimating the parameters of a model that best fits the given data, while ignoring outliers contained in the data due to noise or erroneous measurements. In what follows we report a description of the RANSAC for the detection of vanishing point and the calculation of those lines that originated it (inliers):

- Two lines are randomly selected from the list of lines detected by Hough and the coordinate of their intersection point is calculated. This can be done by solving for $[x, y]^T$ in the following:

  $$\begin{bmatrix} \cos(\theta_1) & \sin(\theta_1) \\ \cos(\theta_2) & \sin(\theta_2) \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} \rho_1 \\ \rho_2 \end{bmatrix}$$

  Where ρ and θ correspond to the parameters for the first line (analogously for the second line). Since there are as many equations as unknowns, the above linear system has only one solution.

- After the coordinates (x,y) of the intersection point are computed, we calculate the distance D of each line to this point by using the following equation:

  $$D_i = x\cos(\theta_i) + y\sin(\theta_i) - \rho_i$$

  Lines with a distance below a certain threshold t are considered inliers, while lines which are more distant are considered outliers. The inlier-set represents the consensus for the vanishing point calculation.

- If the consensus set is larger than a certain threshold T_c, then we consider that these inliers are a good estimate for the inliers of our data and the algorithm stops by returning the estimated intersection point coordinates as well as the inlier set. Otherwise we iterate the previous step either until we do find a large enough consensus or until a maximum number of iterations is reached. In every intermediate iteration with a higher consensus, we keep the new estimate as a better one.

The three thresholds of the RANSAC algorithm are set adaptively.

- The consensus threshold is set as a fraction of the filtered lines considered by RANSAC and has been found to work effectively in the region of 50-80% of the total number of detected lines.
- The maximum number of iterations is set in order to make the algorithm computationally efficient by avoiding the execution for every possible sample. If *p* is the probability that at least one of the samples is free from outliers, *ε* is the probability of a line being an outlier and *s* is the sample size, then it can be proven that: $(1-w^s)^N = 1-p$, where $w = 1-\varepsilon$ [101]. Therefore, the maximum number of iterations can be expressed as $N = \frac{\log(1-p)}{\log(1-(1-\varepsilon)^s)}$,where p is usually set to 99%.
- The distance threshold t is calculated as follows. According to [101], the measurement error can be described by a zero mean Gaussian with standard deviation *σ*. Then the square of the points' distance to the lines can be expressed as a chi-squared distribution with co-dimension equal to 1. Thus, from the

cumulative chi-squared distribution F_m with *m* degrees of freedom (co-dimension):

$$t^2 = F_m^{-1}(a) \cdot \sigma^2$$

Where *α* is the probability of a point being an inlier. Then for m=1 and α=95% :

$$t^2 = 3.84 \cdot \sigma^2 \Rightarrow t = \sqrt{3.84 \cdot \sigma^2}$$

After the completion of the above procedure, we have a number *n* of inlier lines with the highest consensus which we will use to determine the vanishing point in the image. The resulting system will be in the form:

$$\begin{bmatrix} \cos(\theta_1) & \sin(\theta_1) \\ \cos(\theta_2) & \sin(\theta_2) \\ \vdots & \vdots \\ \cos(\theta_n) & \sin(\theta_n) \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} \rho_1 \\ \rho_2 \\ \vdots \\ \rho_n \end{bmatrix}$$

This over-determined problem can be expressed as $\vartheta\chi = \rho$ and solved in a least-squares sense using SVD.

In addition to obtaining the coordinates of the vanishing point for every frame, we store the coordinates of the last 5 frames and return the median of these coordinates. This acts as a buffering mechanism since it suppresses radical changes in the coordinates of the vanishing point due to incorrect estimation in specific frames. The x-coordinate of this filtered vanishing point is then used to calculate the deviation from the principal point of the camera. In our implementation we assume that the principal point is in the middle of the image and that the camera points parallel to the axis of motion of the robot. The deviation, which takes values from 0 to 255 from far left to far right, is then sent to the controller module using UDP datagrams.

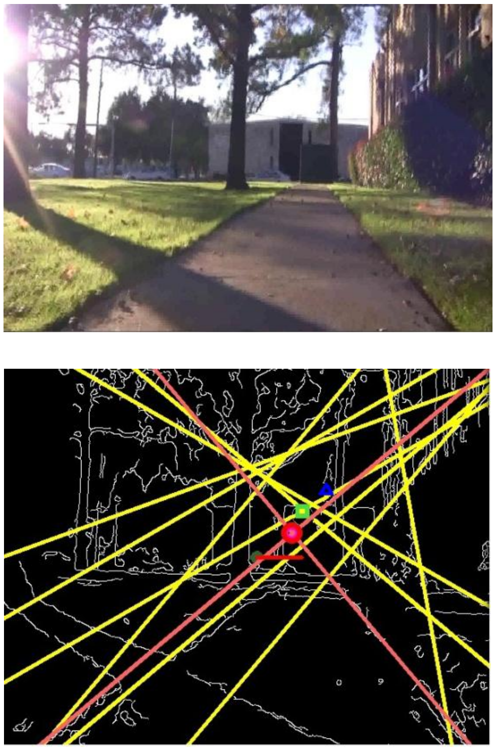


Figure 4-7: Algorithm operation example. Original image (top). Processed edges image with Hough lines, inliers and the estimated vanishing point (bottom).

An example of the operation of the algorithm is seen in Figure 4-7. In Figure 4-7(top) we can see the original image captured by the webcam mounted on the robot. In Figure 4-7(bottom) we can see the image edges with the lines detected from Hough transform. Despite the large number of lines, RANSAC has picked the actual path edges as the inliers (amber lines). The red circle represents the vanishing point estimated for the current frame, the green rectangle is the buffered vanishing point after taking into account the last 5 estimations and the blue triangle is the simple least squares solution. The horizontal red line represents the deviation from the center of the camera and is equal to the x-coordinate of the blue circle. This value is sent to the controller for the steering command calculation.

4.2.3 LIDAR Module

The laser module is designed to serve as the software interface to the LIDAR unit. Measurements from the Hokuyo URG-04-LX-01 ranging unit are acquired using the SKIP 2.0 USB protocol for Hokuyo ranging devices. The URG-04-LX-01 returns readings from an envelope of 240 degrees with 0.36 degree resolution. Each ranging measurement is achieved with a resolution of 1mm, and is assembled into a UDP packet for transmission to the system controller.

Once received by the system controller (at a rate of 10 Hz), the UDP packet is used to populate an array of measurements. These measurements are then filtered using either a moving average or moving median filter, both with adjustable window sizes. After filtering, a “safety envelope” is scanned for potential obstacles. Laser measurements falling within this safety area are counted, and if this number exceeds a user defined threshold, the laser signal overrides the vision signal as the source of error for the PID controller.

4.2.4 System Controller Module

As a central component of the eyeDog, the system controller is responsible for fusing real-time data streams from sensing components. This module performs the selection and filtering of sensor streams in order to generate the control law error signal. At each time step, filtered error signals are used as input to the controller.

The System Control GUI, in addition to conditioning data streams and providing user interface settings, performs discrete iterations of the PID controller. This common controller is used to calculate the turning radius command that is sent to the iRobot platform. The error *e(t)* is defined as the deviation of the vehicle from the desired path at time step *t*. The vehicle turning radius commanded at time *t* is then given by:

$$U(t) = k_p e(t) + k_i \int_{t-T}^{T} e(\tau) d\tau + k_d \frac{de(t)}{dt}$$

The control gains k_p, k_i, and k_d are configurable in the System Control GUI, along with an integration window of size *T*. This allows the user to tune and to evaluate every aspect of the radius controller.

In order to avoid collision with static and moving obstacles, the system controller switches the error signal source from the vision module to the laser module when obstacles are detected within the safety envelope. The centroid of all the scanning points lying within the safety envelope is then used as a new *e(t)* to its upper or lower bound value, such that the platform spins in place away from the obstacle centroid. This behavior is repeated until the safety envelope of the robot is determined to be clear.

The translational velocity command is switched in a manner similar to the turning radius, though a simple bang-bang controller is used in place of PID; its value is set by the user with the control GUI when no obstacles are present (zero otherwise).

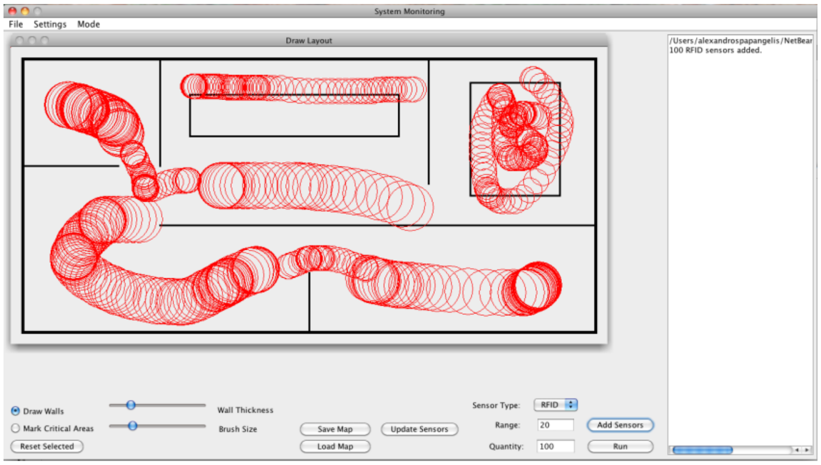
4.2.5 Platform Module

The iRobot platform module receives command packets from the system controller. These command packets are validated and transformed into iRobot Create OI commands, which are sent to the platform through a configurable serial port [102]. In addition to serving as control relay point, the module allows the user to override any UDP command with the keyboard arrows. This feature provides an additional debugging utility for the user.

4.3 Robot Placement Framework

For the placement of the robots, we employed our Sensor Placement and System Monitoring (SPSM) tool [105]. This tool is able to place sensors in optimal positions, according to a map that depicts critical areas and is also able to monitor the system for faulty sensors and raise warnings, alerts or modify the placement if a sensor is faulty. The SPSM tool has two modes, *Environment Drawing* and *System Monitoring* and three phases, defining critical areas, placing sensors and monitoring them afterwards.

In the *Environment Drawing* mode, the user is able to draw a map of the environment (or load a previously saved one), define critical areas with the provided utilities (Figure 4-8) and run the EMSDC [106] algorithm to automatically place sensors in near optimal locations (Figure 4-9). Before running the placement algorithm, however, the user must define the set of available sensors as well as their properties. A sensor may either be “single” or “multiple”, meaning it may sense just one (e.g. sound) or more modalities (e.g. infrared, sound, temperature etc). A sensor also has properties, such as range, coverage angle, battery life and so on. If a sensor is “multiple” the user must define these characteristics for each sub-sensor. However, critical areas need to be defined for each type of sensor, as for example the critical area of a camera is not necessarily the same as the critical area of an RFID reader or a thermometer. Also, a sensor may be considered mobile or static. A static sensor cannot be moved once it is placed, while a mobile sensor, such as our robotic platforms, may move to respond to changes in the environment, depending on the type of mobility (turning, moving, etc). EMSDC also takes into account several other parameters, such as the fact that the more critical an area is the more overlap in coverage we probably want, to increase fault tolerance, or the fact that certain types of sensors may or may not be blocked by obstacles, walls etc. After the tool has placed the sensors, the user may adjust the sensors’ positions to her/is preferences.

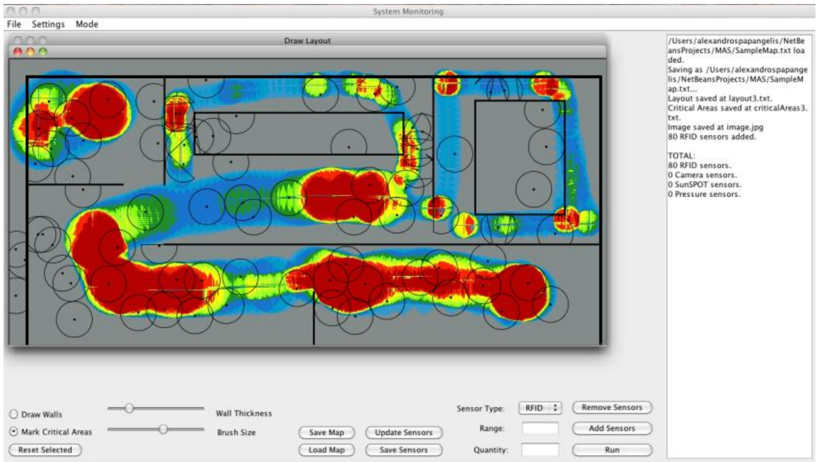


```
/Users/alexandrospapangelis/NetBear
100 RFID sensors added.
```

● Draw Walls
○ Mark Critical Areas

Sensor Type: RFID
Range: 20
Quantity: 100

Figure 4-8: Definition of critical areas in a given environment layout/floor plan using the SPSM tool.



```
/Users/alexandrospapangelis/NetBe
ansProjects/MAS/SampleMap.txt loa
ded.
Saving as /Users/alexandrospapange
lis/NetBeansProjects/MAS/SampleM
ap.txt...
Layout saved at layout3.txt.
Critical Areas saved at criticalAreas3.
txt.
Image saved at image.jpg
80 RFID sensors added.

TOTAL:
80 RFID sensors.
0 Camera sensors.
0 SunSPOT sensors.
0 Pressure sensors.
```

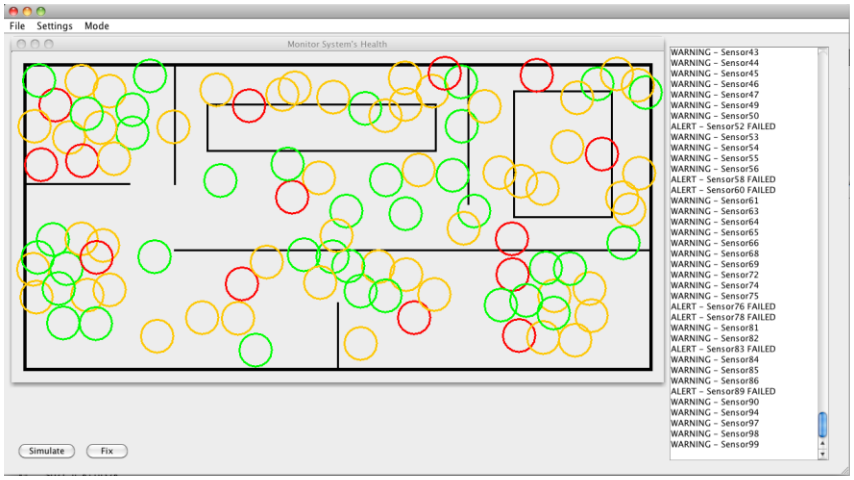
○ Draw Walls
● Mark Critical Areas

Sensor Type: RFID
Range:
Quantity:

Figure 4-9: During sensor placement, in the SPSM tool, sensors move around trying to find an optimal position.

In the *System Monitoring* mode, the system frequently communicates with each sensor to make sure it is working properly. If a sensor fails to communicate within a predefined time frame, a warning is raised. If the sensor does not respond for too long, it is considered failed, an alert is raised and EMSDC will adjust the placement to account for the failure (Figure 4-10).

This tool operates in an online fashion, and is able to adapt to changes in the environment (i.e. changes in the critical areas, sensor failures etc). This gives us the ability to react to events of interest (e.g. the user falls) by simply adjusting the definition of critical areas. E.g. If we have a mobile sensor, it can be moved to compensate for the failure of another sensor or for the appearance of a new critical area.



```
WARNING – Sensor43
WARNING – Sensor44
WARNING – Sensor45
WARNING – Sensor46
WARNING – Sensor47
WARNING – Sensor49
WARNING – Sensor50
ALERT – Sensor52 FAILED
WARNING – Sensor53
WARNING – Sensor54
WARNING – Sensor55
WARNING – Sensor56
ALERT – Sensor58 FAILED
ALERT – Sensor60 FAILED
WARNING – Sensor61
WARNING – Sensor63
WARNING – Sensor64
WARNING – Sensor65
WARNING – Sensor66
WARNING – Sensor68
WARNING – Sensor69
WARNING – Sensor72
WARNING – Sensor74
WARNING – Sensor75
ALERT – Sensor76 FAILED
ALERT – Sensor78 FAILED
WARNING – Sensor81
WARNING – Sensor82
ALERT – Sensor83 FAILED
WARNING – Sensor84
WARNING – Sensor85
WARNING – Sensor86
ALERT – Sensor89 FAILED
WARNING – Sensor90
WARNING – Sensor94
WARNING – Sensor97
WARNING – Sensor98
WARNING – Sensor99
```

Figure 4-10: Monitoring mode of SPSM, where green means the sensor is fine, yellow that there is a warning and red that the sensor is faulty.

4.4 System Operation

As we described in the previous section, the two main system components are the robots and the placement tool, as visualized in Figure 4-11. A typical operation scenario for our system would begin with the definition of the apartment layout by the user. After the layout has been inserted in the tool, the areas of high importance are determined. The importance of each region/room is defined by subjective criteria that usually depend on the time the user spends in a particular area, such as the living room, or increased risk of injury due to the particular characteristics and usage of a specific room, e.g. the restroom. The combination of these important or critical regions create a “critical area map” which is used by the tool as a base map with the default importance values for each area. After these regions, as well as the number of robots available have been defined, the tool uses the EMSDC algorithm to define the number of the robots and position of each one in the different areas of the apartment. During this process, the range and coverage span of each sensor of the robot are taken into account. These factors may depend on the sensitivity of each sensor, its resolution and its directionality. Therefore, the microphone is considered as an isotropic sensor, while the laser sensor has both less effective range as well as a smaller angle of coverage. The camera on the other hand offers good range, but has a very small coverage angle. After all the above parameters are taken into account, the final position of the robots is determined and appropriate commands are sent to each one of them.

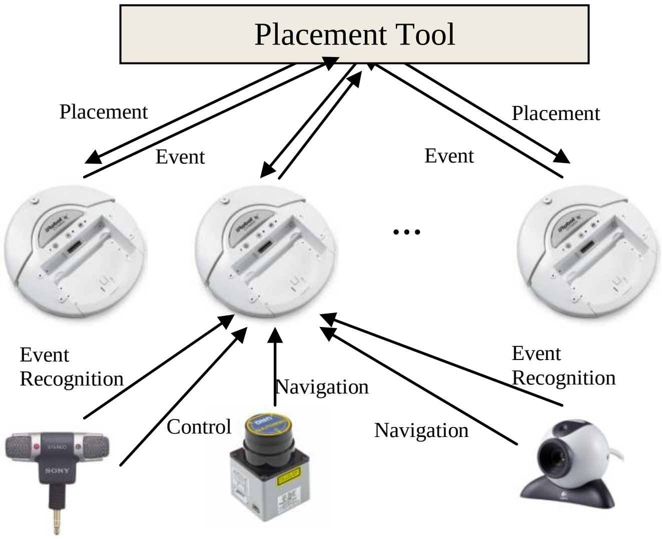


Figure 4-11: System Architecture of the

assistive robot placement framework.

Each robot navigates to its predetermined position by using the camera and the LIDAR sensors. After it reaches its position it starts monitoring the space and stands by for further instructions. Space monitoring could be carried out using both audio and video.

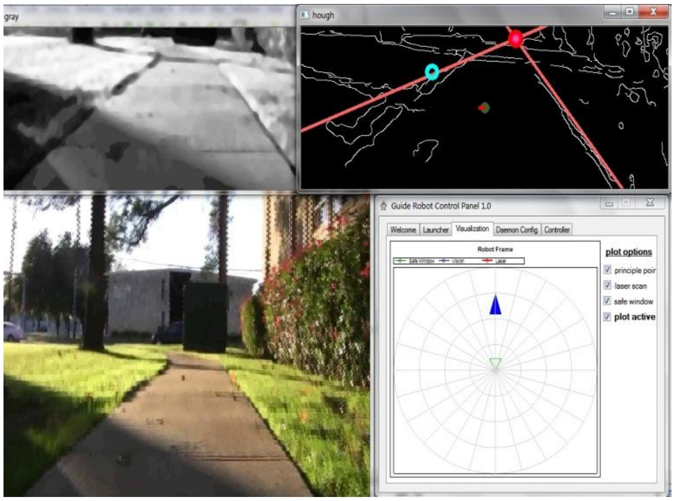
More specifically, the microphone could be used for audio monitoring and the webcam for video monitoring. Audio monitoring can be two-fold, event recognition and vocal commands. In the event recognition mode, the robot would monitor the audio level captured by the microphone. If a high level noise is detected, such as a scream or the sound of an object or person falling, an alert would be sent to the placement tool which raises the importance of the specific area and also records the event type, time and the position where it occurred. The second mode of audio monitoring would use vocal commands to make the robot move. Therefore, the importance map would be updated when a user changes the position of a unit. This mode of operation would give the user the ability to control the robot, but could also allow for further interaction with the robot. The second stream of information which could be used for monitoring is video, based on motion detection. Motion detection conducted by means of background subtraction would be sufficient for a domestic environment with stationary background. In a similar way to audio event recognition, intense motion can be considered a critical event which would cause the placement tool to be updated by increasing the importance of the particular position and record the time, type and position of the event. This can prove very effective in recognizing falls since a concurrent detection of an event from both audio and video can be a good indicator of such an event.

In all cases of an event being detected, the importance of a particular area of the “critical area map” is raised. Nevertheless, in order to avoid the creation of local maxima in the map, the importance of the area is reverted back to its default value found in the base map, after a certain amount of time has elapsed and if no other events have occurred in that area.

4.5 Navigation Experiments

In this section we present the results of the experiments conducted to validate the effectiveness of the robotic design. We tested the system in various scenarios in order to evaluate its robustness to changes in illumination and obstacle avoidance. During our tests we came up with practical modifications that increased both the accuracy and the performance of our system. The first was to restrict the coordinates of the initial intersection point of the pair of lines to the middle 1/3 of the image vertically and inside the boundaries of the image horizontally. This is based on the assumptions that the path cannot have an extreme inclination and that the camera always keeps the path in its view. This made the vanishing point estimation more robust by using RANSAC that discarded false lines intersecting at arbitrary points. The second set of experiments included testing the robotic platform outdoors, as a worst case scenario, to examine the robustness in a radically different environment with greatly varying illumination and contrast conditions as well as fewer perspective lines. The visual module was found to operate better when only the lower half of the image was processed, thus reducing drastically the processing time and the occurrence of false lines coming from objects. The intersection point of the pair of lines was restricted to the upper half of the image.

The highest navigation accuracy was achieved indoors since there was an increased number of perspective lines coming from the walls and the ceiling. This resulted in accurate vanishing point estimation. Outdoors, performance was lower due to the fact that collinear edge pixels coming from objects or noise in the image were detected as lines, introducing inaccuracies in the execution of RANSAC. The collision avoidance system proved to operate robustly in all the environments. The interaction with the robots was done through a GUI from which the system can be switched on and off using two large and distinct buttons. In addition the interface allows for the operating parameters of the robot to be tuned for better performance and for matching the robot's operation to the user's needs (Figure 4-12). Furthermore, the GUI provides a visualization of the captured video with the estimated vanishing point as well as the direction of the robot and any obstacles detected.



plot options

☑ principle poir

☑ laser scan

☑ safe window

☑ plot active

Figure 4-12: Vision and ranging data visualized by the system at different processing steps.

4.6 Conclusions

Our novel framework for optimal assistive robot placement manages to effectively monitor an ambient intelligence environment by taking into account the apartment layout, the importance of different rooms and the user's preferences. It does so by utilizing the EMSDC algorithm, which enables robots incorporating a variety of different sensors to be positioned optimally in an apartment for event recognition. Our system is based on the eyeDog guide robot. This robot was designed as an effective assistive solution for the visually impaired, able to successfully navigate on a given path and avoid obstacles. Additionally, it is cost effective and can be easily built since it uses low cost, off-the-shelf parts. Our system can prove very valuable in an assistive environment for the elderly or disabled were accidents and unexpected events must be detected promptly in order to insure the person’s well-being.

Chapter 5

Discussion and Extensions

We have presented our vision and efforts towards a novel ambient intelligence environment that utilizes speech and localization information as well as robotics. The individual developed subsystems satisfy successfully the main conditions that we set in Chapter 1. More specifically, they exhibit the ability to recognize speech and be location aware, enabling recognition of events and ultimately context-awareness. Furthermore, the incorporation of embedded sensors such as the Kinect and the utilization of depth images instead of planar video, renders them less intrusive, reducing privacy concerns. Finally, they are adaptive to sensor failure or dynamic changes of the environment over time by means of optimal placement of mobile sensor bearing robotic platforms. Therefore, the framework described in this work enables the creation of the infrastructure for effective human-centered computing to provide enhanced pervasive services to humans. Our efforts were successful at focusing on the collection and analysis of multimodal data that can result from human monitoring applications inside an ambient intelligence environment.

In particular, in Chapter 2 we presented a novel multimodal ASR system that utilizes facial depth information captured by the Kinect, in addition to the traditional audio and planar video modalities. We also used this configuration to capture a connected digit database in two languages, the first featuring this stream combination. Feature extraction was employed by means of the 2D DCT and a two-stage LDA feature selection scheme was applied to the visual and depth features in order to boost lip-reading performance. Finally, state synchronous HMMs were used for data fusion and speech modeling. We tested our system under the influence of babble audio noise and 4 types of video degradations and conducted experiments not only in English but also in Greek. Our experimental results demonstrated that the depth modality improves word accuracy in comparison to audio-only and audiovisual recognition and that 3D visual information from both planar video and the depth stream, leads to a significant increase in accuracy in comparison to audio-only ASR.

In Chapter 3 we presented a novel system capable of accurate and robust person localization and identification. Our system combines the tracking capabilities of the Kinect sensor with identification information from existing RFID technology. 2 Kinect devices were used as well as 2 RFID antennas, identifying and tracking multiple tags. The three types of data captured to solve the localization problem were the RSSI, 3D depth images and audio information. Accurate position estimation for each person was carried out using the depth sensor of the Kinect, by means of skeletal tracking. The system was deployed in a simulated apartment and during the experimental phase it was tested for a 4 person scenario. The results achieved exhibited high identification and localization accuracy, exceeding 90%, proving its effectiveness.

Finally, in Chapter 4 we expanded on the concept of ubiquitous computing in a dynamic ever-changing environment by presenting a framework for adaptive monitoring through optimal placement of sensor bearing robots. This framework can ensure optimal coverage of the monitored environment as well as adaptation and failure recovery in an online fashion by means of the Extended Max Sum Decentralized Coordination algorithm. The foundation of this framework is the prototype eyeDog robotic platform initially developed as a guidance aid for the visually impaired. The main components of this platform are a camera for the navigation task, and a LIDAR unit for obstacle avoidance. RANSAC has been employed for vanishing point estimation based navigation using a set of adaptive thresholds over similar approaches, and a PID controller ensures stable steering control. The feasibility and effectiveness of the approach was demonstrated by successful navigation down the center of a path, while navigating away from obstacles and optimal sensor placement.

5.1 Extensions

A number of possible extensions and improvements can be considered for each one of the subsystems described in this work. Our AVASR system demonstrated the usefulness of 3D visual information for this task. However, its overall robustness could benefit through the use of product-modal HMMs, which lack the inherent assumption of synchronous streams. In addition, investigating the extraction of a different set of features from the visual and particularly the depth stream, could lead to higher accuracy by ensuring extraction of higher information content from this type of data. Finally, in order to promote natural interaction of the user with the environment, a dialogue system as well as speech synthesis would be a meaningful addition for a potential real world deployment of the system.

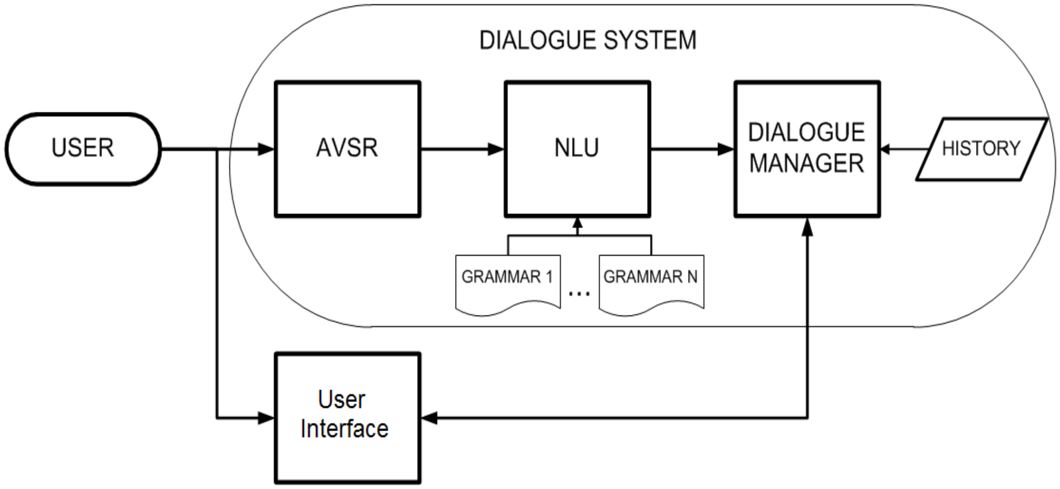


Figure 5-1: Proposed extension with the incorporation of a dialogue system that will enable effective interaction between the user, the ambient intelligence environment and the robotic platforms.

In terms of our multimodal person localization and identification system, there is plenty of room for future extensions and improvements. In the localization domain, we consider as a reasonable and straight forward expansion to experiment with stereovision based on multiple cameras or Kinect sensors as an additional information stream. We also consider experimenting with the integration of an advanced data fusion technique for combining the heterogeneous data from all sensors and especially from audio and video. After confirming the effectiveness of this design, it could be extended by utilizing depth and audio information from additional Kinect devices for increased robustness and coverage.

In terms of our robotic infrastructure, an extension regarding the SPSM tool would be to automatically generate patrol routes for the mobile sensors and adjust them accordingly if events of interest occur. Another prospective development would be to replace the webcam with the MS Kinect sensor in order to capture real-time video with depth information from the robot’s surroundings. This would allow us to acquire depth measurements that are correlated with the video stream and also replace both the webcam and the LIDAR unit with only the Kinect. Furthermore, if it proved sensible, we could disable the planar video input for increased privacy. Another foreseeable enhancement would be to add a speech based user interface for ease of use, more tailored towards the elderly or disabled. By combining our ASR system with the robotic platform and also incorporating speech generation and a dialogue system as part of a prompting system, it would be possible to prompt the user in case an event is detected and also allow her/him to issue commands verbally and receive system notifications. A final possibility is the development of a different type of platform, e.g. a quad rotor such as the ARdrone, in order to examine different challenges for the navigation task and promote easier access to otherwise inaccessible portions of a domestic environment by our current platform.

References

[1] L. Rudolph, “Project Oxygen: pervasive, human-centric computing an initial experience", In Proc. of the International Conference on Advanced Information Systems Engineering, pp. 1-12, 2001.

[2] A. Jaimes, D. Gatica-Perez, N. Sebe and T. S. Huang, “Guest editors' introduction: Human-Centered Computing Toward a human revolution”, Computer, vol. 40, num. 5, pages 30-34, 2007.

[3] M. Weiser, “The computer for the 21st century”, Scientific American vol. 265, num. 3, pages 94-104, 1991.

[4] E. Zelkha, B. Epstein, S. Birrell and C. Dodsworth, “From Devices to "Ambient Intelligence"”, In Proc. of the Digital Living Room Conference, vol. 6, June 1998

[5] K. Park, L. Yong, V. Metsis, Z. Le and F. Makedon, “Abnormal human behavioral pattern detection in assisted living environments”, In Proc. of the 3rd International Conference on Pervasive Technologies Related to Assistive Environments (PETRA), pp. 9-17, ACM, 2010.

[6] E. Aarts, R. Harwig and M. Schuurmans, "Ambient Intelligence" chapter in “The Invisible Future: The Seamless Integration of Technology Into Everyday Life”, McGraw-Hill Companies, 2001

[7] B. Schilit, N. Adams and R. Want, “Context-aware computing applications”, In Proc. of the IEEE Workshop on Mobile Computing Systems and Applications (WMCSA), pp. 89–101, 1994.

[8] B.N. Schilit and M.M. Theimer, “Disseminating Active Map Information to Mobile Hosts”, IEEE Network, vol. 8 num. 5, pages 22–32, 1994.

[9] A. K. Dey, “Understanding and Using Context”, Personal Ubiquitous Computing vol. 5 num. 1, pages 4–7, 2001.

[10] T. Gu, Z. Wu, X. Tao, H. K. Pung and J. Lu, “epSICAR: An Emerging Patterns based Approach to Sequential, Interleaved and Concurrent Activity Recognition”, In Proc. of the 7th Annual IEEE International Conference on Pervasive Computing and Communications (Percom), pp. 1 - 9, 2009.

[11] K. Iwano, S. Tamura and S. Furui, “Bimodal speech recognition using lip movement measured by optical-flow analysis”, In Proc. of HSC, pp.187-190, 2001.

[12] S. Nakamura, H. Ito and K. Shikano, “Stream weight optimization of speech and lip image sequence for audio-visual speech recognition”, In Proc. of the International *Conference* on Spoken Language Processing (ICSLP), vol. 3, pp. 20-24, 2000.

[13] G. Potamianos, C. Neti, G. Gravier, A. Garg and A.W. Senior, “Recent advances in the automatic recognition of audio-visual speech”, Invited, In Proceedings of the IEEE, vol. 91, num. 9, pp. 1306-1326, 2003.

[14] E.K. Patterson, S. Gurbuz, Z. Tufekci and J. N. Gowdy, “CUAVE: A new audio-visual database for multimodal human-computer interface research”, In Proc. of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP), vol. 2, pp. 2017-2020, 2002.

[15] J.R. Movellan, “Visual speech recognition with stochastic networks”, In Advances in Neutral Information Processing Systems, D. Toruetzky, G. Tesauro and T. Leen, Eds., vol. 7. MIT Press, Cambridge, 1995.

[16] C. C. Chibelushi, S. Gandon, J. S. D. Mason, F. Deravi and R. D. Johnston, “Design issues for a digital audio-visual integrated database”, In Proc. of IEEE Colloquium on Integrated Audio-Visual Processing for Recognition, Synthesis and Communication, pp. 1–7, 1996.

[17] K. Messer, J. Matas, J. Kittler, J. Luettin and G. Maitre, “Xm2vtsdb: The extended m2vts database”, In Proc. of the International Conference on Audio and Video-based Biometric Person Authentication, pp. 965-966, 1999.

[18] R. Goecke and B. Millar, “The audio-video Australian English speech data corpus AVOZES”, In Proc. of the International Conference on Spoken Language Processing (ICSLP), vol. 3, pp. 2525–2528, 2004.

[19] R. Goecke, “3D lip tracking and co-inertia analysis for improved robustness of audio-video automatic speech recognition”, In Proc. of the International Conference on Auditory-Visual Speech Processing (AVSP), pp. 109-114, 2005.

[20] A. Vorwerk, X. Wang, D. Kolossa, S. Zeiler and R. Orglmeister, “WAPUSK20 - A database for robust audiovisual speech recognition”, In Proc. of the International Conference on Language Resources and Evaluation (LREC), 2010.

[21] A. Ortega, F. Sukno, E. Lleida, A. Frangi, A. Miguel, L. Buera and E. Zacur, “AV@CAR: A Spanish multichannel multimodal corpus for in-vehicle automatic audio-visual speech recognition”, In Proc. of the International Conference on Language Resources and Evaluation (LREC), vol. 3, pp. 763–767, 2004.

[22] 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. of the IEEE Aerospace Conference, pp. 10, 2006.

[23] M. Zamalloa, L.J. Rodriguez, M. Penagarikano, G. Bordel and J.P. Uribe, “Comparing genetic algorithms to principal component analysis and linear discriminant analysis in reducing feature dimensionality for speaker recognition”, In Proc. of the Genetic and Evolutionary Computation *Conference* (GECCO), pp. 1153-1154, 2008.

[24] G. Potamianos and H.P. Graf, “Linear discriminant analysis for speechreading”, In Proc. of the IEEE International Workshop on Multimedia Signal Processing (MMSP), pp. 221-226, 1998.

[25] G. Potamianos and C. Neti, “Improved ROI and within frame discriminant features for lipreading”, In Proc. of the IEEE International Conference on Image Processing (ICIP), vol. 3, pp. 250-253, 2001.

[26] G. Galatas, G. Potamianos, D. Kosmopoulos, C. McMurrough and F. Makedon, “Bilingual corpus for AVASR using multiple sensors and depth information”, In Proc. of the International Conference on Auditory-Visual Speech Processing (AVSP), pp. 103-106, 2011.

[27] G. Galatas, G. Potamianos and F. Makedon, “Audio-visual speech recognition incorporating facial depth information captured by the Kinect”, In Proc. of the European Signal Processing Conference (EUSIPCO), pp. 2714-2717, 2012.

[28] G. Potamianos, H.P. Graf and E. Cosatto, “An image transform approach for HMM based automatic lipreading”, In Proc. of the IEEE International Conference on Image Processing (ICIP), vol. 3, pp. 173-177, 1998.

[29] G. Galatas, G. Potamianos and F. Makedon, “Audio-visual speech recognition using depth information from the Kinect in noisy video conditions”, In Proc. of the International Conference on Pervasive Technologies Related to Assistive Environments (PETRA), pp. 2-6, 2012.

[30] The Primesensor Reference Design, [Online] Available at: http://www.primesensor.com

[31] G. Bradski and A. Kaehler, “Learning OpenCV: Computer vision with the OpenCV library”, O’Reilly Media, 1st edition, September 2008.

[32] C. M. Bishop, “Pattern Recognition and Machine Learning”, Springer, Heidelberg, 2006.

[33] J. Shain, C. B. Owen and F. Makedon, “Detecting lip motion in digital video”, In Proc. of SPIE Multimedia Systems and Applications, vol. 3528, pp.15-25, 1999.

[34] G. Potamianos and P. Scanlon, “Exploiting lower face symmetry in appearance-based automatic speechreading”, In Proc. of the International Conference on Auditory-Visual Speech Processing (AVSP), pp. 79-84, 2005.

[35] 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.

[36] “The HMM-based speech synthesis system (HTS)”, [Online] Available at: http://hts.sp.nitech.ac.jp

[37] G. Galatas, G. Potamianos and F. Makedon. “Robust multimodal speech recognition in two languages utilizing video and distance information from the Kinect”, In Proc. of the International conference on Human-Computer Interaction, pp. 43-48, 2013.

[38] A. Varga and H. Steeneken, “Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems”, Speech Communication, vol. 12, num. 3, pages 247-251, 1993.

[39] G. Galatas, G. Potamianos, A. Papangelis and F. Makedon, “Audio visual speech recognition in noisy visual environments”, In Proc. of the International Conference on Pervasive Technologies Related to Assistive Environments (PETRA), pp. 19-23, 2011.

[40] 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, vol. 82, num. 8, 1050–6, 2001.

[41] J. A. Stevens, “Fatalities and injuries from falls among older adults”, MMWR, vol. 55, num. 45, 2006.

[42] H. Nait-Charif and S. J. McKenna, “Activity summarization and fall detection in a supportive home environment”, In Proc. of the International Conference on Pattern Recognition (ICPR), vol.4, pp. 323- 326, 2004.

[43] C. Rougier, J. Meunier, A. St-Arnaud and J. Rousseau, “Fall Detection from Human Shape and Motion History Using Video Surveillance”, In Proc. of the Advanced Information Networking and Applications - Workshops (AINAW), vol.2, pp. 875-880, 2007.

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

[45] T. Teixeira, D. Jung and A. Savvides, “Tasking networked CCTV cameras and mobile phones to identify and localize multiple people”, In Proc. *of the ACM International Joint* Conference *on Pervasive and* Ubiquitous Computing *(Ubicomp), pp.* 213–222, 2010.

[46] R. Cucchiara, M. Fornaciari, A. Prati and P. Santinelli, “Mutual calibration of camera motes and RFIDs for people localization and identification”, In Proc. of the ACM/IEEE International Conference on Distributed Smart Cameras (ICDSC), 2010.

[47] C. H. Kuo and R. Nevatia, “How does Person Identity Recognition Help Multi-Person Tracking?”, In Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1217-1224, 2011.

[48] D. Mitzel, E. Horbert, A. Ess and B. Leibe, “Multi-person tracking with sparse detection and continuous segmentation”, In Proc. of the European Conference on Computer Vision (ECCV), pp. 397-410, 2010.

[49] A. Ess, B. Leibe, K. Schindler and L. V. Gool, “Robust Multi-Person Tracking from a Mobile Platform”, In IEEE Transactions on Pattern Analysis and Machine Learning (PAMI), vol. 31, 2009, pp. 1831–1846.

[50] M. Andriluka, S. Roth and B. Schiele, “People Tracking-by-Detection and People Detection by Tracking”, In Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1-8, 2008.

[51] B. Wu and R. Nevatia, “Detection and Tracking of Multiple, Partially Occluded Humans by Bayesian Combination of Edgelet Part Detectors”, International Journal of Computer Vision (IJCV), pages 247–266, 2007.

[52] D. Gavrila and S. Munder,“Multi-Cue Pedestrian Detection and Tracking from a Moving Vehicle,” International Journal of Computer Vision (IJCV) vol. 73, pages 41–59, 2007.

[53] J. Krumm, S. Harris, B. Meyers, B. Brumitt, M. H. and S. Shafer, “Multi-camera multi-person tracking for easy living”, In Proc. of the International Workshop on Visual Surveillance (IWVS), pp. 3-10, 2000.

[54] D. Focken and R. Stiefelhagen, “Towards Vision-Based 3-D People Tracking in a Smart Room”, In Proc. of the IEEE International Conference on Multimodal Interaction (ICMI), pp. 400-405, 2002.

[55] Y. Chang, R. Yan, D. Chen and J. Yang, “People identification with limited labels in privacy-protected video”, In Proc. of the IEEE International Conference on Multimedia and Expo (ICME), pp. 1005-1008, 2006.

[56] B.S. Ashar and A. Ferriter, "Radiofrequency identification technology in health care: benefits and potential risks", JAMA: The Journal of the American Medical Association, vol. 298, pages 2305-2307, Nov. 2007.

[57] K. Finkenzeller, “RFID handbook: fundamentals and applications in contactless smart cards and identification”, Wiley, New York, NY, USA, 2003.

[58] T. J. Hazen, E. Weinstein, B. Heisele, A. Park and J. Ming, “Multimodal Face and Speaker Identification for Mobile Devices”, Book Chapter in Face Biometrics for Personal Identification: Multi-Sensory Multimodal Systems, R. I. Hammoud, B. R. Abidi, and M. A. Abidi (Eds.), pp. 123-138, Springer, 2007.

[59] N. Fox, R. Gross, P. Chazal, J. Cohn, and R. Reilly, “Person Identification Using Automatic Integration of Speech, Lip and Face Experts,” In ACM Special Interest Group on Multimedia (SIGMM) Multimedia Biometrics Methods and Applications, pp. 25-32, 2003.

[60] H. K. Ekenel, M. Fischer, Q. Jin and R. Stiefelhagen, “Multimodal person identification in a smart environment”, In Proc of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Biometrics Workshop, USA, 2007.

[61] T. Rappaport, “Wireless Communications: Principles and Practice,” Prentice Hall PTR, Upper Saddle River, NJ, USA, 2001.

[62] D. Hahnel, W. Burgard, D. Fox, K. Fishikin and M. Philipose, “Mapping and localization with RFID technology”, In Proc. of the IEEE International Conference on Robotics and Automation (ICRA), pp.1015-1020, 2004.

[63] M. Bouet and A. L. D. Santos, “RFID tags: Position principles and localization techniques”, In Proc. of Wireless Day, pp. 1-5, 2008.

[64] L. Ni, Y. Liu , Y. Cho Lau and A. Patil, “LANDMARC: Indoor location sensing using active RFID”, Wireless Networks, pages 701–710, 2004.

[65] P. Wilson, D. Prashanth and H. Aghajan, “Utilizing RFID signaling scheme for localization of stationary objects and speed estimation of mobile objects”, In Proc. of the IEEE International Conference on RFID, pp. 94-99, 2007.

[66] D. Hahnel, W. Burgard, D. Fox, K. Fishkin and M. Philipose, “Mapping and localization with RFID Technology”, TR IRS-TR-03-014, Intel Research, 2003.

[67] S. Ferdous, G. Galatas, K. Vyas, E. Becker, L. Fegaras and F. Makedon. “Multi-person identification and localization in pervasive assistive environments”, IJCSMR vol. 4, no. 1, pages 751-758, 2012.

[68] G. Galatas, S. Ferdous and F. Makedon, “Multimodal person localization and emergency detection using the Kinect”, International Journal of Advanced Research in Artificial Intelligence, vol. 2 num. 1 pages 41-46, 2013.

[69] G. Galatas and F. Makedon, “A system for multimodal context-awareness”, International Journal of Advanced Computer Science and Applications, vol. 4, num. 9, pages 130-136, 2013

[70] G. Galatas, S. Ferdous and F. Makedon. “Multi-person Identification and Localization for Ambient Assistive Living”, In Proc. of Distributed, Ambient, and Pervasive Interactions- Conference on Human-Computer Interaction, pp. 109-114., 2013.

[71] J. Hightower, C. Vakili, G. Borriello and R. Want, “Design and Calibration of the SpotON Ad-Hoc Location Sensing System”, unpublished, Seattle, WA, August 2001.

[72] D. Joho,C. Plagemann and W. Burgard, “Modeling RFID signal strength and tag detection for localization and mapping”, In Proc. of the IEEE International Conference on Robotics and Automation (ICRA), pp. 3160-3165, 2009.

[73] K. Bernardin and R. Stiefelhagen, “Audio-visual multi-person tracking and identification for smart environments”, In Proc. of the 15th international conference on Multimedia, pp. 661-670 2007.

[74] A. Salah, R. Morros, J. Luque, C. Segura, J. Hernando, O. Ambekar, B. Schouten and E. Pauwels, “Multimodal identification and localization of users in a smart environment”, In Journal on Multimodal User Interfaces (JMUI), vol. 2, pp. 75–91, 2008.

[75] M. M. Trivedi, K. S. Huang and I. Mikic, “Dynamic context capture and distributed video arrays for intelligent spaces”, IEEE Transactions on Systems: Man and Cybernetics - Part A: Systems and Humans, vol. 35, January 2005.

[76] L. Klingbeil and T. Wark, “A wireless sensor network for real-time indoor localization and motion monitoring”, In Proc. of the International Conference on Information Processing in Sensor Networks, pp. 39-50, 2008.

[77] J. A. Stankovic, Q. Cao, T. Doan, L. Fang, Z. He, R. Kiran, S. Lin, S. Son, R. Stoleru and A. Wood, “Wireless sensor networks for in-home healthcare: potential and challenges”, In Proc. of the joint workshop on High Confidence Medical Devices, Software, and Systems (HCMDSS), pp. 2-3, 2005.

[78] L. Zhang and Z. Wang, “Integration of RFID into wireless sensor networks: Architectures, opportunities”, In Proc. of the International Conference on Grid and Cooperative Computin Workshops (GCCW), pp. 463–469, 2006.

[79] S. Rowe and C. Wagner, “An Introduction to the Joint Architecture for Unmanned Systems (JAUS)”, Technical Report, Cybernet Systems Corporation, *Ann Arbor* 1001 (2008): 48108.

[80] G. Marrocco, “The art of UHF RFID antenna design: impedance-matching and size-reduction techniques”, IEEE Antennas and Propagation Magazine, vol. 50, pp. 66–79, 2008.

[81] G. Galatas, A. Papangelis and F. Makedon. “A framework for optimal assistive robot placement for event recognition”, In Proc. of IEEE International conference on Computing, Networking and Communications (ICNC) – Cyberphysical Systems Workshop, pp. 200-204, 2013.

[82] G. Galatas, C. McMurrough, G. Mariottini and F. Makedon. “eyeDog: an assistive-guide robot for the visually impaired”, In Proc. of 4th International Conference on Pervasive Technologies Related to Assistive Environments (PETRA), p. 58. ACM, 2011.

[83] “The Microsoft Kinect SDK”, [Online] available at: http://msdn.microsoft.com/en-us/library/hh855347.aspx

[84] “The Microsoft Kinect Skeleton tracker”, [Online] available at: http://msdn.microsoft.com/en-us/library/jj131025.aspx

[85] S.P. Levine, D.A. Bell, L.A. Jaros, R.C. Simpson, Y. Koren and J. Borenstein, “The NavChair Assistive Wheelchair Navigation System”, IEEE Transactions on Rehabilitation Engineering, vol.7, no.4, pages 443-451, Dec, 1999.

[86] “Guide Dogs of America - FAQ”, [Online] available at: http://www.guidedogsofamerica.org/1/mission/

[87] “DARPA Grand Challenge”, [Online] available at: http://www.darpa.mil/grandchallenge/index.asp

[88] "Mini Grand Challenge Robot Contest", [Online] available at: http://www.cede.psu.edu/users/avanzato/robots/contests/outdoor/index.htm

[89] K. Macek, B. Williams and S. Kolski, "A lane detection vision module for driver assistance", In Proc. of the IEEE/APS Conference on Mechatronics and Robotics, 2004.

[90] Y. Malinovskiy, Y. Wu and Y. Wang, "Video-Based Vehicle Detection and Tracking Using Spatiotemporal Maps", In Proc. of the 88th Annual Transportation Research Board Meeting, pp 81-89, 2009.

[91] Zu Whan Kim, "Robust Lane Detection and Tracking in Challenging Scenarios", IEEE Transactions on Intelligent Transportation Systems, vol. 9, num. 1, pages 16-26, March, 2008.

[92] A. Lopez, C. Canero, J. Serrat, J. Saludes, F. Lumbreras and T. Graf, "Detection of Lane Markings based on Ridgeness and RANSAC," In Proc. of the IEEE Conference on Intelligent Transportation Systems, pp 733-738, 2005.

[93] M. Aly, "Real time detection of lane markers in urban streets," In Proc. of the IEEE Intelligent Vehicles Symposium, pp 7–12, 2008.

[94] M. A. Fischler and R. C. Bolles, "Random sample consensus: A paradigm for model fitting with application to image analysis and automated cartography", Communications of the ACM, vol. 24 num. 6, pages 381-395, 1981.

[95] V. Kulyukin, C. Gharpure and J. Nicholson, "RoboCart: Toward Robot-Assisted Navigation of Grocery Stores by the Visually Impaired" In Proc. of the IEEE/RSJ International Conference on Intelligent Robots and Systems pp. 2845-2850, 2005.

[96] V. Kulyukin, C. Gharpure, J. Nicholson and G. Osborne, “Robot-Assisted Wayfinding for the Visually Impaired in Structured Indoor Environments”, Autonomous Robots, vol. 21 num. 1, pages 29-41, June 2006.

[97] V. Kulyukin and G. Gharpure, “A Robotic Shopping Assistant for the Blind”, In Proc. of the 29th Annual Conference of the Rehabilitation Engineering and Assistive Technology Society of North America (RESNA), 2006.

[98] G. Rezai-Rad and H.H. Larijani, “A New Investigation on Edge detection Filters Operation for Feature Extraction under Histogram Equalization Effect” In Proc. of Geometric Modeling and Imaging, pp 149-153, 2007.

[99] K. Wong, “Canny Edge Detection Auto Thresholding”, [Online] Available at: http://www.kerrywong.com/2009/05/07/canny-edge-detection-auto-thresholding, May, 2009.

[100] R.C. Gonzalez and R.E.Woods, “Digital Image Processing”, 3rd Edition. Prentice Hall, NJ. p. 2001.

[101] R. Hartley and A. Zisserman, “Multiple View Geometry in Computer Vision”, Second Edition, Cambridge University Press, March, 2004.

[102] “Multiple iRobot Create Open Interface (OI) specification”, iRobot Corporation, [Online] Available at: http://www.irobot.com, 2006.

[103] “The OpenCV library”, [Online] Available at: http://opencv.willowgarage.com/wiki/

[104] M. A. Patricio, J. Carbo, O. Perez, J. Garcia and J. M. Molina, “Multi-agent framework in visual sensor networks”, EURASIP Journal on Advances in Signal Processing, 2007.

[105] A. Papangelis and F. Makedon, “A tool for sensor placement and system monitoring in assistive environments” In Proc. of the International Conference on Pervasive Technologies Related to Assistive Environments (PETRA), 2011.

[106] A. Papangelis, V. Metsis, J. Shawe-Taylor and F. Makedon, “Sensor placement and coordination via distributed multi-agent cooperative control”, In Proc. of International Conference on Pervasive Technologies Related to Assistive Environments (PETRA), 2010.

Biographical Information

Georgios Galatas received his PhD in Computer Engineering in 2013 from the Computer Science and Engineering department at the University of Texas at Arlington. During his doctoral studies he was 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 at the University of Patras in 2008. His research interests include Computer and Robot Vision, Digital Image Processing and Automatic Speech Recognition. He has co-authored several peer reviewed papers published in technical conferences and journals.