ISTANBUL TECHNICAL UNIVERSITY ★ GRADUATE SCHOOL OF SCIENCE ENGINEERING A

ACTIVE AUDIO-VISUAL HUMAN TRACKING FOR ROBOTS

M.Sc. THESIS

Barış BAYRAM

Department of Computer Engineering

Computer Engineering Programme

MAY 2015

ISTANBUL TECHNICAL UNIVERSITY ★ GRADUATE SCHOOL OF SCIENCE ENGINEERING A

ACTIVE AUDIO-VISUAL HUMAN TRACKING FOR ROBOTS

M.Sc. THESIS

Barış BAYRAM
(504131507)

Department of Computer Engineering

Computer Engineering Programme

Thesis Advisor: Asst. Prof. Dr. Gökhan İNCE

MAY 2015

İSTANBUL TEKNİK ÜNİVERSİTESİ ★ FEN BİLİMLERİ ENSTİTÜSÜ

ROBOTLAR İÇİN AKTİF İŞİTSEL-GÖRSEL İNSAN TAKİBİ

YÜKSEK LİSANS TEZİ

Barış BAYRAM
(504131507)

Bilgisayar Mühendisliği Anabilim Dalı

Bilgisayar Mühendisliği Programı

Tez Danışmanı: Asst. Prof. Dr. Gökhan İNCE

MAYIS 2015

Barış BAYRAM, a M.Sc. student of ITU Graduate School of Science Engineering and Technology 504131507 successfully defended the thesis entitled **"ACTIVE AUDIO-VISUAL HUMAN TRACKING FOR ROBOTS"**, which he prepared after fulfilling the requirements specified in the associated legislations, before the jury whose signatures are below.

Thesis Advisor :	**Asst. Prof. Dr. Gökhan İNCE** Istanbul Technical University	.....
Jury Members :	**Asst. Prof. Dr. Sanem SARIEL** Istanbul Technical University	.....
	Asst. Prof. Dr. Sinan KALKAN Middle East Technical University	.....
		.....
		.....

Date of Submission : **04 May 2015**
Date of Defense : **22 May 2015**

FOREWORD

Acknowledgment is given to Dr. Gökhan İnce for his assitance and time, and to the Istanbul Technical University Research Fund under Grant 39537 for supporting the project. He provided excellent facilities and advices during this project.

May 2015

Barış BAYRAM
(MSc. Student)

TABLE OF CONTENTS

Page

ABBREVIATIONS

SSL	**:** Sound Source Localization
ROS	**:** Robot Operating System
OpenCV	**:** Open Source Computer Vision
HARK	**:** Honda Research Institute Japan Audition for Robots with Kyoto University
FFT	**:** Fast Fourier Transform
PCL	**:** Point Cloud Library
Camshift	**:** Continuously Adaptive Meanshift
fps	**:** frame per second
GEVD	**:** Generalized EigenValue Decomposition
MUSIC	**:** MUltiple SIgnal Classification
DoA	**:** Directio of Arrival
CM	**:** Correlation Matrix

LIST OF TABLES

Page

LIST OF FIGURES

Page

ACTIVE AUDIO-VISUAL HUMAN TRACKING FOR ROBOTS

SUMMARY

In the field of robot perception, sensing solutions in robot audition, object detection, object tracking, and navigational tasks are developed in order to understand the environment, and intelligent robot behaviors based on the solutions are achieved to provide better and natural interaction with humans. Moreover, methods for integration of the different kinds of perception abilities like audition and vision, by considering the solutions for auditory and visual tasks are investigated and developed by inspiring from human tendency in better understanding the environment and obtaining more reliable results than understanding only with a single ability.

In this thesis, a multimodal system is designed in the form of an active audio-visual perception in order to improve the perceptual capability of a robot in a noisy environment. The system running in real-time uses 1) audition modality, 2) a complementary vision modality and 3) motion modality incorporating intelligent behaviors based on the data obtained from both sensory modalities. In this system, the detection, localization and tracking of the speaker are the main tasks of audition and vision modalities independently.

Multiple signal classification based on generalized eigenvalue decomposition method is utilized for sound source localization in audition modality; and a method combining feature-based and color-based face detection and tracking methods are proposed in vision modality to cope with issues affecting the detection and tracking in a real-world environment.

The goal of the motion modality is to enable a robot having intelligent and human-like behaviors by using the results obtained from the sensor fusion framework, which is more reliable and robust than the results obtained from unimodal systems. This system is tested on a real robot and the performance of sensory modalities in both single and multi-person experiments are evaluated. The contribution of sensor fusion in the tracking task is confirmed to be improved compared to the performances of each one of the sensory modalities.

Obstacle detection and avoidance without losing the person as another perceptive ability and intelligent behaviors of the robot tracking and approaching the human is investigated and implemented. The performances in robot perception and execution of behaviors are observed and tested in a real environment having obstacles.

ROBOTLAR İÇİN AKTİF İŞİTSEL-GÖRSEL İNSAN TAKİBİ

ÖZET

Robot algılama alanında, çevreyi işitsel ve görsel olarak anlamak için, robot işitmesinde, nesne tespiti ve takibinde, ve navigasyonel görevlerde algılama çözümleri geliştirilmektedir. Bu çözümlere göre daha iyi ve doğal bir etkileşim sağlamak için robot davranışları tasarlanmaktadır. İnsanların çevreyi daha iyi anlamak ve daha güvenilir sonuçlar elde etmek için kullandıkları eğilimden etkilenilerek, işitsel ve görsel algı yetilerinin birleştirilerek kullanıldığı yöntemler de incelenmekte ve geliştirilmektedir.

Bu tezde, işitsel ve/veya görsel olarak karmaşık çevreler ve şartlar altında robotların algılama yetisini iyileştirmek için, çoklu kipli/sensörlü sistem tasarlanılmıştır. Gerçek zamanda koşan bu sistem 1) işitme kipi, bu kipe destek olması amacıyla 2) görme kipi ve bu iki kipten gelen verilere dayalı akıllı davranış örüntüleri yaratmak için önerilen 3) hareket kipi olmak üzere üç kipten oluşmaktadır. Ortamdaki konuşmacının tespiti, lokalizasyonu ve takibi, işitme ve görü kiplerinin, bağımsız olarak gerçekleştirdiği başlıca görevleridir.

İşitme kipinin görevlerinde, Genelleştirilmiş Özdeğer Ayrıştırma (Generalized EigenValue Decomposition - GEVD) tabanlı Çoklu Sinyal Sınıflandırması (MUltiple SIgnal Classification - MUSIC) ile ses kaynağı lokalizasyonu metodu kullanılmıştır. Görme kipinde, gerçek bir çevrede bulunan, tespiti ve takibi etkileyen sorunlarla mücadele edebilmek için sunulan öznitelik ve renk tabanlı yüz tespiti ve takibi yöntemlerinin birleştirildiği bir sistem kullanılmıştır. Bu yöntemde, gereken önişlemlerden sonra Haar öznitelikleri tabanlı yüz alanı tespiti uygulanır. Ardından, tespit edilen alanın yüz olduğu konusundaki kesinliği arttırmak için, bu alan üzerinde Haar öznitelikler tabanlı en az bir göz aranır. Kesinliği biraz daha arttırmak için yüz ve göz tespiti yapılan alandaki ten rengi oranları hesaplanır. Daha sonra, bu oranların tanımlanmış eşik değerlerle karşılaştırılmasının yapıldığı bir sınırlama uygulanmaktadır. Bu sınırlamadan da geçen tespit sonuçları yüz olarak kabul edilerek, posizyon bilgileri Camshift (Continuously Adaptive Meanshift) renk tabanlı takip algoritmasına gönderilmiştir. Ten rengi kontrolü ve takip algoritması için gerekli olan ten rengi ayıştırmassı uyarlanabilir YCbCr renk uzayı kullanılarak gerçekleştirilmektedir. Uyarlanabilirlik, kişinin veya robotun ani hareketlerinden kaynaklı, ışıktaki beklenmedik değişimler, renk tabanlı uygulamaları eetkileyeceği için ihtiyaç duyulmuş ve renk uzayının alt sınırları, tespit edilen yüz alanı dışındaki ten rengi oranlarına bakılarak gerçekleştirilmektedir. Aynı zamanda, tespit edilen alan üzerinde, eigenface tabanlı yüz tanıma uygulanır.

Hareket kipinin amacı ise robotlara, işitme ve görü kiplerinden gelen konum bilgilerinin birleştirilmesi ile elde edilen daha güçlü ve güvenilir konum bilgisini

kullanarak insan benzeri hareketler gerçekleştirmektir. Bu birleştirme parçacık filtrelemesi kullanılarak, takip edilen kişinin gelecekteki posizyon bilgisi tahmin edilerek gerçekleştirilmiştir. Aynı zamanda, engel sakınma esnasında robotun hareketlerindeki yer değişimleri bu filtrelemede kullanılmıştır. Çünkü, bu yer değişimleri sebebiyle takip edilen kişinin kaybedilme ihtimali vardır. Bu sebeple, hesaplanan yer değişimlerine göre, robotun kişinin nerede lduğu hatırlaması amacıyla filtrelemeye dahil edilmiştir. Her bir sensörlü kipin performansı ve önerilen sistemin performansı gezgin bir robot platformu üzerinde gerçek dünyada tek kişi ve çoklu kişiler içeren deneyler ile değerlendirilmiştir. Sensör tümleştirmesinin katkısı, insan takibi performansının her iki kipe göre de iyileştiği teyit edilmiştir.

Bu çalışmada, robot kazandırılan akıllı davranış örüntüleri, sensör füzyonu sonucuna göre insan takibi ve insana yaklaşmayı içeren yakınlık tabanlı yaklaşım hareketleri ve engel tespitine ve posizyonuna göre füzyon sonucundan bağımsız olarak takip edilen hedefin yerini kaybetmeden gerçekleştirilen engel sakınma hareketleridir. Bu yaklaşım hareketleri, yüz tanınması sonucuna göre belirgin bir kişiye karşı da uygulanabilmektedir. Engel sakınma hareketleri, tespit edilen engelin kapsadığı alana göre, uygun olan tarafa doğru robotun geçebileceği genişlikte bir yol bulunana kadar döngüsel hareketleri ve bulunduktan sonra belli bir süre doğrusal hareketi içerir. Bu sakınma hareketlerinde gerçekleştirilen yer değişimleri de füzyon sonucuna dahil edilerek, engelden kurtulduktan sonra, hedefin olduğu tarafa doğru dönmesi ve onu tekrar bulması hedeflenmiştir. Robot algısının ve davranışlarının performansı, engelli gerçek bir ortamda gözlemlenmiş ve test edilmiştir.

Gerçekleştirilen deneylerde, farklı ışık şartlarına sahip 3 farklı kayıt ile görü kipinin yüz tespiti ve takibindeki başarımı ve sensör füzyonunun ses kaynağı lokalizasyon sonucuna katkısı, aynı anda ve ayrı ayrı konuşan iki kişi ile hareket halindeki bir konuşmacı kayıtları üzerinde değerlendirilmiştir. Aynı zamanda, iki konuşmacı arasında yüz tanıma ile belirli bir konuşmacıdan gelen işitsel-görsel veriler üzerinde sensör füzyonu uygulanması da test edilmiştir. Akıllı davranış örüntüleri ise; robotun hareket halindeki bir konuşmacı ve hareketsiz iki konuşmacı ile etkileşimi değerlendirilmiş ve engelli bir ortamda hareketsiz bir konuşmacı ile etkileşimi üzerinde gözlemlenmiştir.

1. INTRODUCTION

The importance and the necessity of robot perception in realistic environments with obstacles -especially while the robots are moving- increase in everyday environments, where the robots are deployed. In social and home environments, the robots equipped with this perception ability and the intelligent behavior patterns can be utilized as a peer and a helper for people, who are especially old, disabled, and having problems in supplying their needs. Moreover, they can be utilized in security and rescue missions as well as surveillance and patrol duties. For these reasons, the main requirement of the robot is to detect and to enhance the continuity of tracking the humans, who need to interact with robots.

Autonomous navigational tasks in real environment with obstacles are considerable challenges in robotics. The tasks are achieved by mobile robots used for research, assisting humans, military operations, cleanning the room, mapping and other missions without damaging the environments, humans and themselves. Therefore, obstacle detection and avoidance abilities are required to accomplish the tasks efficiently.

1.1 Purpose of Thesis

The aims of the thesis are to design a multimodal system having three modalities as audition, vision and motion in order to detect and to track humans and to implement the system with a real robot in real environments with obstacles to design intelligent behaviors due to the results of the sensory modalities. Using the audition modality, sound source localization is achieved to find the position of powerful waveform data from a speaker. In the vision modality, the main tasks are to detect and to track the speaker visually to be fused by using a sensor fusion framework based on particle filter tracking technique and to detect obstacles. The fusion results and information about the obstacles are transmitted to the motion modality to enable the robot behaving due to the information in order to interact with the speakers. In addition, the multimodal

system is able to cope with unreliable data from a sensor, and to improve the reliability of robot decisions by integrating the data from complementary sensors.

This system is applied to multi-person experiments, and the multimodal system working in realistic environments is tested with a mobile robot and humans. It is observed that a multi-sensor system is more effective in tracking than any kind of unimodal system.

1.2 Literature Review

The literature overview of this research covers a variety of aspects mainly from four domains: 1) Audio processing, 2) Computer vision, 3) Audiovisual integration and 4) Active motion.

Audio Processing: Sound processing methods implemented on robots provide natural and comfortable human-robot interaction [1] by inspiring from humans communication, because humans use environmental sounds such as speech sounds to interact with each other. Sound Source Localization (SSL), sound source separation and speech recognition are used for human-robot interaction [2, 3]. Nakadai et al. proposed the application of the techniques in the robot audition in dynamically-changing acoustic environments where the speakers and/or the robots are moving. However, only tracking a single sound source is may be insufficient because of discontinuity in speech signals due to silent pauses and also noises more dominant than the useful sound sources [4]. Kim proposed techniques for robots in order to select a desired speaker in multiple detected sound sources and faces to track for interaction in daily life environments. In this research, a probability based method is developed for auditory and visual information integration, and for production of reliable path in real-time. Moreover, Murray [5] et al. poposed a method for sound source localization and tracking by using Interaural Time Difference (ITD), and cross-correlation and auditory cues to compute the angle of a sound-source on the horizontal plane.

Computer Vision: Visual data processing as an alternative to sound based interaction may also experience difficulties under a few situations like abnormal lighting conditions, occlusions on the target objects, going out of the eyesight, etc. In the

literature, face detection is the primary component of human tracking. Color-based face detection methods can be inadequate on unexpected changes in lights, in case of similar color of skin and background, and existence of different kinds of skin-colors [4]. On the other hand, feature-based face detection methods like Haar cascade classifier can fail because of distance to the camera, light conditions and similarity of features between objects and a face [4]. For feature based face detection, Viola and Jones [6] proposed a method to rapidly detect the facial features in an image to achieve high detection rates by using integral image representation for computing Haar features quickly, Adaboost learning algorithm to select a number of important features from a large dataset on the image to yield efficient classifiers, and cascading step to combine multiple Adaboost classifiers for quickly discarding the background regions of the image. Shen et al. used Haar features extraction proposed by Viola and Jones, to retrieve similar faces by using a predefined face database [7]. Moreover, by using the cascade classifier method, facial features detection like eyes, lips and noise and regionalization on the search area to improve the accuracy of the detection are described in [8]. In color based face detection methods, several color spaces are used such as RGB, YCbCr and HSI [9]. Singh et al. compared these three color spaces on skin color extraction and proposed a method combining the spaces to obtain more accurate extraction. A hybrid face detection is commonly utilized to increase the accuracy by combining feature based and color based face detection methods. Niazi and Jafar [10] proposed a hybrid method searching Haar features on skin color extracted areas by using HSV color space, and Maghraby et al. [11] described a hybrid method combining three detectors that one is for detecting near-frontal upper-bodies by using Haar cascade classifier, the second one is for face detection by using the same classifier and the last one is for skin color extraction on the region of interest where the second detector does not detect any faces. Wang and Tan [12] described a method for face detection based on shape information on the images with simple background by using an energy function to link the extracted edges to finally extract a face contour. Moreover, deformable templates can be used for feature extraction from faces. Yuille et al. [13] proposed a facial feature like eyes extraction method by using the templates according to an energy function to define which links, edges and peaks intensity to the properties of the eye template. For human tracking, a computer vision system

is designed called as Pfinder by detecting the boundaries of a person and analyzing the inside of the boundaries [14]. Huang, Gutta, and Wechsler [15] proposed a novel algorithm for face detection by using decision trees (DT).

The color-based tracking algorithms are utilized to track an object such as Meanshift [16] and Camshift (Continuously Adaptive Meanshift) [17]. In the [16], Cheng proposed the tracking algorithm by finding the peak(mode) values of probability distributions of colors to track the object through the values. Likewise, Bradski described the tracking algorithm Camshift by improving the Meanshift to adapt the moving object by dynamically changing the probability distributions. Jacquin and Eleftheriadis [18] described a method for tracking a localized face and facial features by detecting head boundary and identifying eyes-noise-mouth features inside the boundary on binary images in head-and-shoulders video sequences. In addition, Hua et al. [19] proposed a fast approach for multiple faces detection and tracking in real-time by building color models to capture the inherent chrominance of the skin color with some additional morphological processing and filtering to delineate facial regions. To determine the identity of the detected face, several rapid face recognition methods [20] are developed. Turk et al. proposed a rapid technique by searching the identity of a detected face into a set of face images by comparing eigenvectors of the correlation matrix of the images to find the best fit. Before the face detection, recognition and tracking algorithms, several preprocessing techniques are applied to improve the accuracy of the results such as morphological operations, histogram equalization and filtering [21], [22] required for removing the outliers to have only an object tracked efficiently by the color-based tracking algorithms, and for improving the quality of images.

Audio-visual Integration: Audio-Visual (AV) integration can be utilized in many research fields of robotics, e.g. to improve the noise-robustness of voice activity detection [23] and speech decoding [24]. Yoshida described on audio-visual integration making automatic speech recognition more robust and reliable against the distance of a speaker in the environments having acoustically noisy or multiple speech sounds, and presented two-layered method covering Audio-Visual Voice Activity Detection (AV-VAD) integrating several AV features by using a Bayesian network in

order to robustly detect voice activity or the duration of speech, and Audio-Visual Speech Recognition(AVSR) estimating the reliability of acoustic features and visual features due to the results of AV-VAD layer. On the other hand, Koiwa described two approaches to improve the robustness of AVSR, which one is AV integration approach for coping missing feature in auditory and visual features and the other is a biologically-inspired approach grouping phoneme and viseme assumed as auditory and visual units for coarse-to-fine recognition. Likewise, multimodal human detection and tracking gives more efficient and reliable results [25] because tracking with unimodal systems is too difficult due to the visual and auditory problems in a real environment. Nakamura et al. proposed robust speaker tracking based on intelligent SSL for a robot in real environment. In the intelligent tracking process [25], three issues are coped as robustness against powerful noises, lack of a general framework for selective listening to sound sources, and tracking of inactive and/or noisy sound sources by utilizing extended localization method, gaussian mixture model based sound source identification and particle filter based audio-visual integration, respectively. Pavlovic [26] proposed a new multimodal framework for audio-visual feature prediction and classification based on Hidden Markov Models(HMMs). Moreover, Wang and Brandstein [12] described a hybrid face tracker based on audio-visual data, acoustically estimating the positions of the speakers from microphone array data while precise localization and tracking are derived from visual data in real-time.

Active Motion: Active audition, which integrates audition, vision and motor control improves the quality and the performance of auditory processing using active motion. In the robot audition community, active audition has been applied to sound source localization. Turning motion is the essential motor action in some of these studies [4], [27], [28], [29]. An active audition system proposed in the early 2000's [27] controlled the upper torso of a robot to align two microphones towards a sound source in order to localize and to track sound sources in given multiple sound sources scenes. Rodemann et. al described an active SSL system using binaural and spectral cues, which utilizes pan-and-tilt motion to track objects in azimuth and elevation positions [28] by using three microphone where two of them is used for estimating azimuth angle of sound source, and the other microphone orthogonal is required for

estimation of elevation of the sources. Berglund also proposed an active audition framework for SSL [29] by steering two microphones orthogonal to a sound source which also makes use of reinforcement learning techniques. Kim proposed a system comprising sound source localization, voice activity detection and face/sound source tracking by using motor command to trigger turning behavior [4]. Although all these techniques demonstrated high localization performances, they controlled just the body part that the microphones were mounted on. Some researches proposed to use locomotion of the robot, namely the displacement of the body in 3D world [30], [31]. Sasaki proposed to utilize sound source mapping based on triangulation [30]. However, they mainly focused on improving the localization algorithm and did not consider the path planning aspects. Martinson et al. proposed a robot, which goes further away from different noise sources to improve the signal to noise ratio [31]. However, they used active audition mainly as a tool to enhance a text-to-speech application, not to improve the friendliness and quality of human robot interaction. Moreover, by using Honda ASIMO and Humanoid SIG2 with 8-ch microphone arrays, Yamamoto et al. [32] described a robot audition system consisting of sound source localization and separation with a microphone array, and system integration based on missing feature theory (MFT) to cope with noises in real-time for automatic speech recognition by using two and three simultaneous speech signals.

1.3 Outline of the Thesis

In the rest of the thesis, a multimodal human tracking system for active audio-vision covering robot perception and active robot behaviors is explained, and the technical details of each sensory modality and sensor fusion framework are described in Chapter 2. The experiments for the multimodal system, the proposed face detection method, the sensor fusion framework and robot behaviors are elaborated in Chapter 3. The last Chapter concludes the thesis. Appendix A provides information about the command-line commands required during the implementation of the project. Appendix B presents the network and modules used for sound source localization. Appendix C includes the data flows in the implementation of the multimodal system as a graph.

2. MULTIMODAL SYSTEM FOR ACTIVE AUDIO-VISION

In this chapter, the multi-modal system including the robot perception and the active robot behaviors based on the perception abilities are explained in briefly.

The multimodal system (Fig. 2.1) designed in this work consists of three modalities, namely audition, vision and motion. The proposed system works as follows: The audition modality performs sound source localization on the auditory data streams captured by microphones and vision modality performs the proposed face detection and tracking method, and obstacle detection procedure on the visual data streams captured by two cameras. The localization results about a person interacting with the robot using both sensory modalities are integrated by a sensor fusion framework and then transmitted to the motion modality for converting this data into action by using the motors of a robot, and the avoidance actions are achieved independently from the sensor fusion through the output of obstacle detection. The movements causing the loss of the target are considered to be integrated in the sensor fusion framework. Therefore, the robot interacts with the person in this continuous loop comprising the perception and action. The technical details will be explained in the consequent chapters.

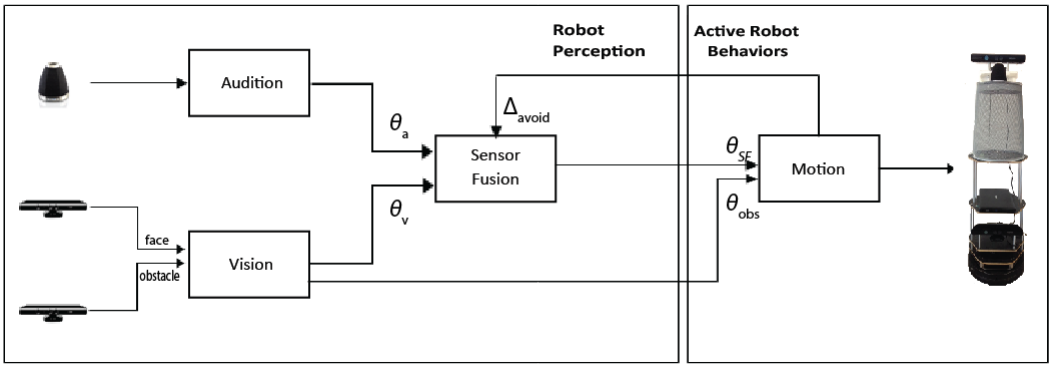

Figure 2.1: An overview of the system architecture.

2.1 Robot Perception

To understand the environment and to find solutions for auditory and visual problems as two sensory modalities in this research. The aims explained in this chapter are *detection* and *localization* of powerful sound sources in the environments in the framework of the audition modality, and *face detection and tracking*, and *obstacle detection* in the framework of the vision modality.

2.1.1 Audition modality

Speaker localization is important in the application of speech enhancement techniques which require information of the speaker's exact position. Based on this position, the microphone array may be steered towards the found direction for more effective sound and speech acquisition. This method is adequate for speech enhancement applications where a moving speaker exists, such as in video conference calls or robot audition. A localization system not only can be applied to single talker scenarios, but may also be used in a multi-speaker case to enhance speech of a particular speaker with respect to others' utterances or with respect to interfering noise sources.

By using a noise-robust method, the sound source localization is achieved to be integrated with the visual position of the speaker. Thus, audition is the first modality which starts the interaction with human and determines the position of the speaker for the vision modality.

2.1.1.1 GEVD-based sound source localization

In the audition modality, a noise-robust sound source localization method, MUltiple SIgnal Classification (MUSIC) based on Generalized EigenValue Decomposition (GEVD) [25] is utilized.

This localization process starts with converting audio signals from the time domain to the frequency domain by applying Fast Fourier Transform(FFT). The audio signals

captured by each microphone are expressed as;

$$x_i(t) = \sum_{j=1}^{L} a_{i,j}(t,\theta_j)s_j(t,\theta_j) + n_j(t), \tag{2.1}$$

where t denotes time, i is the index of microphones, M is the number of microphone, j is the index of sound sources, L is the number of sound sources, θ_j represents the direction of the j^{th} sources, s_j is the signal in the time domain, $a_{i,j}$ is the transfer function between j^{th} sound source and i^{th} microphone in the time domain and n_m denotes the additive noise representing noises in the environment.

The FFT of the noisy signal is designated as;

$$\boldsymbol{X}(\omega) = \sum_{j=1}^{L} \boldsymbol{A}_j(\omega,\theta_j)\boldsymbol{S}_j(\omega,\theta_j) + \boldsymbol{N}(\omega), \tag{2.2}$$

where ω denotes the frequency, $\boldsymbol{S}_j$ is the signal in the frequency domain and $\boldsymbol{A}_j$ is the transfer function vector for each microphone, haviing transfer functions, $A_{i,j}$ between j^{th} sound source and i^{th} microphone in the frequency domain, and $\boldsymbol{N}$ is the additive noise vector where each noise is measured by a microphone.

In FFT, the waveform data is analyzed with a frame length, L designated as;

$$L = \frac{f_s x}{1000}, \tag{2.3}$$

where f_s [Hz] denotes the sampling frequency data, and x [ms] is temporal length of a window.

To determine steering vector, $\boldsymbol{G}(\omega,\psi)$ of waveform data derived by using FFT and ψ is the azimuth of a sound source from a multichannel microphone array as impulse responses. The responses are converted with the predefined transfer functions used for propagating the sounds from impulses;

$$\boldsymbol{G}(\omega,\psi) = \boldsymbol{A}_j(\omega,\theta_j). \tag{2.4}$$

GEVD-MUSIC is applied on the FFT output sampled from the audio signals to find the number of sound sources and to estimate the directions of arrival of each sound source by performing eigenvalue decomposition on the correlation matrix (CM) of the overall noisy signal, by separating subspaces of undesired interfering sources as well

as the sound sources of interest, and eventually by determining the peaks occurring in the spatial spectrum. A successive source tracker performs a temporal integration for a given time window.

Firstly, the CM, $\boldsymbol{C}(\omega)$ of the output from FFT is estimated in this GEVD-MUSIC as a sound source localization method;

$$\boldsymbol{C}(\omega) = \boldsymbol{X}(\omega)\boldsymbol{X}^{T}(\omega), \tag{2.5}$$

where ^{T} denotes the transpose operator.

and then, the eigenvalue decomposition is achieved as;

$$\begin{aligned} \boldsymbol{C}(\omega)\boldsymbol{e}_i(\omega) &= \boldsymbol{\lambda}_i(\omega)\boldsymbol{K}(\omega)\boldsymbol{e}_i(\omega), \qquad 1 \leq i \leq M, \\ \boldsymbol{K}^{-1}(\omega)\boldsymbol{C}(\omega)\boldsymbol{e}_i(\omega) &= \boldsymbol{\lambda}_i(\omega)\boldsymbol{e}_i(\omega), \qquad 1 \leq i \leq M, \end{aligned} \tag{2.6}$$

where $\boldsymbol{e}_i$ represents eigenvalues in the frequency domain, $\boldsymbol{\lambda}_i$ is a scalar value for eigenvalues and $\boldsymbol{K}(\omega)$ denotes the configurable CM utilized to extend standard eigenvector decomposition as GEVD. The $\boldsymbol{K}(\omega)$ is configured as;

$$\boldsymbol{K}(\omega) = \boldsymbol{N}(\omega)\boldsymbol{N}^{T}(\omega), \tag{2.7}$$

where the correlation matrix is computed by using $\boldsymbol{N}(\omega)$, the additive noise vector. Therefore, the configurable CM whitens the noisy sound source by aligning the eigenvalues, $\boldsymbol{e}(\omega)$.

After the extended decomposition of eigenvalues, the spatial spectrum for estimation of DoA (Direction of Arrival) of the sound sources is computed as;

$$\boldsymbol{P}(\theta) = \frac{|\boldsymbol{G}^{\boldsymbol{T}}(\omega,\theta)\boldsymbol{G}(\omega,\theta)|}{\sum_{i=L+1}^{M} |\boldsymbol{G}^{\boldsymbol{T}}(\omega,\theta)\boldsymbol{e}_i(\omega)|}, \tag{2.8}$$

where the DoA estimated (Fig. 2.2) is θ for L peaks of the spatial spectrum.

Because of aligning the eigenvalues with respect to the target/non-target sources, the azimuth values of the sources are correlated with the power of each source in order to select eigenvalue by assuming that the power of all sound sources are larger than the noises. Therefore, the noisy sound sources are separated from the target sources.

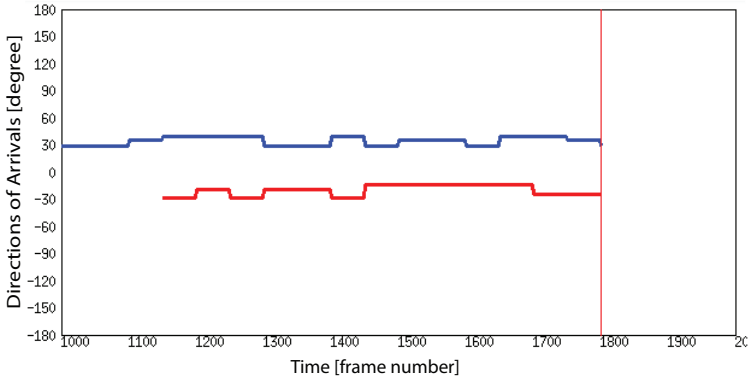

Figure 2.2: The estimated DoA of the detected two sound sources.

To decide on the sound sources for localization is achieved by comparing the power of peaks with a predefined threshold value. The sources having the power less than the threshold are eliminated.

2.1.2 Vision modality

In real-world environments, the visual problems affecting the color and feature based face detection are unstable illumination, occlusion of face with other objects, turning faces and eyesight out of camera. To cope with these problems, a method (Fig. 2.3) is proposed, which involves feature-based face as well as eyes detection by using a learning algorithm and applies a constraint on the color of the detected area by using a color space to extract skin colors so that more accurate and robust speaker detection is achieved by eliminating the false positives.

After the precision on a detected face is provided, face recognition step is applied, and then the face whose identity is known is tracked by using an object tracking algorithm. During this visual perception process for face detection, the color space affecting the result of the constraint and the performances of other color based techniques is changed dynamically according to the amount of skin color distribution in the facial area being detected and tracked, and in the entire area not having the facial area.

To detect obstacles while approaching a detected face by the proposed method, a virtual laser scan is created through converting point clouds by finding the closest points between all points of each column in the clouds.

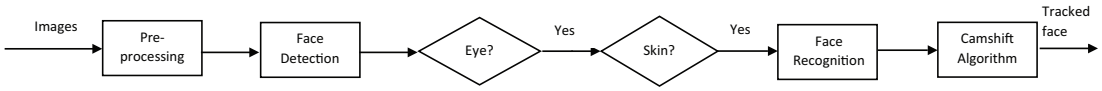

Figure 2.3: The proposed method for face detection and tracking in the vision modality.

2.1.2.1 Preprocessing

Preprocessing step used in this proposed method before feature detection and skin color extraction is required to improve accuracy of the detection and extraction by removing the outliers in images and normalizing the image colors in order to increase the quality of the images.

Before feature based detection training a classifier on a gray-scaled image, RGB images are converted to the gray-scaled images, and then, histogram equalization technique is utilized on the gray-scale image to adjust the contrast on the image by using its histogram values.

Skin color extraction based on a color space causes lots of outliers distributed on the image. After extraction considering the range of the color space, a binary image having outliers is created through the space, where white pixels represent the color desired to be extracted, which is the skin color for face detection. Therefore, morphological operations [21], [22] are required for removing the outliers to have only the areas including a great amount of pixels having the white color. Firstly, the erosion operation is applied twice to discard the useless skin color pixels by removing the pixels on the boundaries, and then adding white pixels on the boundaries of area having the color by applying dilation operation twice. It is important how many times the operations will be used not to affect negatively the amount of white pixels.

In Fig. 2.4, the morphological operators are shown on a color extracted binary image in (a), and then the outliers are removed by using erosion twice in (b). To fill the gaps created because of the erosion, dilation is performed twice to the image for achieving more accurate color extraction output as in (c).

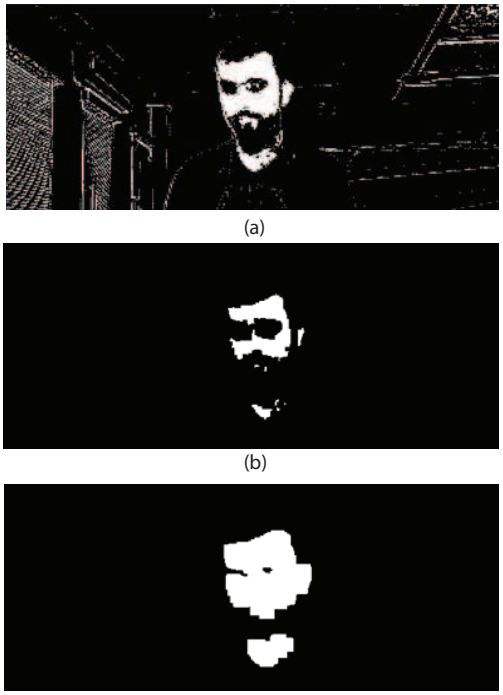

Figure 2.4: Morphological operations results.

2.1.2.2 Haar cascade classifier

Haar-features are calculated on an image by using Adaboost cascade classifier [6] as the starting step of the proposed face detection method. The Haar-features use contrast variances in adjacent rectangular regions on a searching area on an image and a facial Haar-feature is formed by two or more adjacent regions with the contrast variances.

The rapid facial feature detection technique is considered to be more efficient in finding faces in a real noisy environment, even if it detects lots of false-positives. The other advantage of using these features is to be easily scaled by changing the regions' size. Therefore, the detection is adaptable on various sizes.

After gray-scaled image conversion and applying histogram equalization on it, facial Haar-features are searched on the image, and then for each detected face, features of eyes are searched on the face. The search parameters need to be adjusted to small values without diminishing the possibility to detect a face far away from the camera. Therefore, Haar cascade classifier detects lots of objects as face and also having skin-color. The eye detection decreases the probability of the false positive instances. Likewise, the reliability of eye detection is increased by regionalizing the searching area with the detected face. The regionalization technique is applied on the

only these areas according to a specified scale factor used to incrementally scale the detection resolution, which are less than faces.

Selecting a small scale factor for eyes detection does not affect the time complexity of algorithm too much because of the regionalization, although it is important for face detection to be used on the entire frame. The number of detection results are increased when the scale factors and the size of searching area for face and eyes features are decreased. Therefore, the appropriate values should be selected to run in real-time.

To improve the precision of eye detection, two constraints on the location of eyes are utilized. These are to check the location of a detected eye whether or not it is on the upper side of a face, and relation between the two detected eyes.

2.1.2.3 Skin color constraint

This feature based detection step is also not sufficient to ensure the high detection rates. Thus, we added a constraint on the detected faces with skin color information by using an adaptive YCbCr color space.

To improve the accuracy of constraint results and the performances of color based methods, the boundaries of YCbCr color space for skin color are adaptively changed when unexpected changes in illumination occur.

The color space is composed of three components, Y, Cb and Cr where Y is the luminance component computed from RGB and Cb and Cr are the blue and red differences from the luminance component as chroma components. The lower threshold of Cr is the one what is changed adaptively. The pixels on an image are analyzed and a binary image is created where the pixels having the RGB color values between the threshold values of Cb and Cr are set to be white and the rest of the pixels are set to be black.

The detected facial area including eyes are checked whether or not it has a predefined amount of pixels having a skin color RGB values by using the color space to be certain whether the area is a face. The area is converted to a binary image through the color space and then the ratio of white pixels is computed to be compared with a threshold value which is close to the mean of the proportions on the previous detection results,

updated after each face detection. The detected feature based areas are counted as having skin if they have the rate which is bigger than the threshold value.

2.1.2.4 Camshift algorithm

There are still a few problems in a real-world environment; namely the room may have unstable light conditions, the speaker may not look at the camera directly, or he/she may be far away from the robot affecting the detection of eyes. As a consequence, the visual method may also eliminate some true positives. To deal with these problems, the position of a detected face area is sent to the color-based tracking algorithm, Camshift [17] to provide continuity in visual tracking.

The tracking algorithm based on the color distribution is computationally efficient for real environments and it is required to deal with the inadequacy of Haar cascade classifier in detecting faces not frontal. After skin color extraction and morphological preprocessing operations, Camshift calculates probability of the skin color distribution on a given search region and tracked window.

The performance of Camshift is not based only on the color distribution, but also the region and the tracked window location on the image. Due to some illumination problems and the color of objects not face, being in the color space in a real environment, the image may have areas affecting the result of Camshift. For these problems, size of the region and the location and size of tracked window should be determined carefully.

An initial tracked window is required to start Camshift algorithm in each frame. To specify its location and size is essential not to lose the person or not to capture all white pixels not belonging to a face when the person and/or the robot is moving. Therefore, the search region and tracked window should be determined as including the face and considering velocity of the movements for each time. The width and height displacements of the tracked face according to the previous frame due to the person or robot movements are estimated and the means of width and height displacements are computed to be used as the width and height value of the region.

Another important parameter is the location of search windows on the image. The position of the tracked face on the previous frame is used as the center point of the search window for the next frame.

The main difference from the other common algorithm for color-based tracking, Meanshift is that the size of capturing window is not stable and is adaptively changed due to the changes in the face sizes as Camshift [33]. The size of a face tracked is essential for localization and proximity estimation. By using Meanshift, if the face is detected on the position far away from the camera, and tracking process is started from that face, the window capturing the face will remain stable even the person is approaching the camera. Thus, it may lead to the unreliable localization results which are the center point of the stable window and inadequate proximity estimation according to the depth information in that window.

This adaptivity in YCbCr color space is required in order to make more robust tracking in Camshift and to decrease the number of eliminated true positives because of light conditions. The lights may change when a person moves, and enlarges the active area of Camshift to cover light changes. Therefore, the position of the detected area can be shifted towards the wrong direction.

The proposed face and eyes detection with skin-color constraint is still applied in conjunction with the Camshift algorithm because also this color-based tracking algorithm has problems such as going out of the eye-sight and unexpected changes in light. Therefore, if any face is not detected on the tracking area in a while, Camshift algorithm is paused until a detection occured because this area focused by Camshift may consist of an object having skin-color, but not a face.

2.1.2.5 Face recognition

The recognition step is applied if a face is detected for the first time. When working in real-world environments and with a real robot. the rapid recognition technique which is based on eigenfaces is utilized because quickly recognizing the detected faces is an important requirement for low time-complexity. The eigenfaces are created with a number of sample faces and their different poses by computing eigenvectors of the covariance matrix of set of the faces to be stored.

2.1.2.6 Proximity estimation

The proximity to the detected person is calculated to create intelligent motion patterns for robots as further explained. By utilizing RGB-D data sensed with infrared light sensor in Kinect camera, the distances of the objects to the camera are predicted. Each pixels of facial area on the RGB depth raw image is examined to determine how close it is to the camera (Fig. 2.5) shows the coloring scheme according to the object's closeness blue meaning too far from the camera and red meaning too close, or no coloring if it is on the desired position. After the coloring operation, the color distributions on the area are computed in order to measure the proximity to the robot.

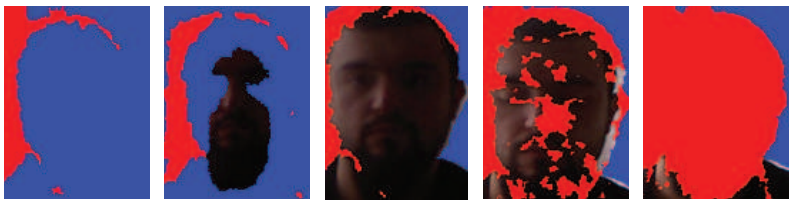

Figure 2.5: Depth information in order to compute proximity while approaching to the person. The rate of blue or red colored pixels show the distance to the camera according to their depth values. The high rate of blue ones like in (a) and (b) shows too low proximity, and the high rate of red ones like in (d) and (e) shows too high proximity. In (c), the high rate of uncolored pixels show the desired proximity.

2.1.2.7 Obstacle detection

The obstacle detection is an important ability for a moving robot in the environments having obstacles and to plan trajectory by avoiding them. Point clouds sensed by Kinect equipped on the robot to be close to the floor are examined to use the camera as a virtual laser scan in order to detect an obstacle, to computer its size and to find the best available path. The 3D clouds representing ranges are virtually converted to 2D laser scan as fake. For this conversion, the closest point is found for each column in the cloud by computing range and angle values of each point, and comparing the results with prespecified minimum range, maximum range, minimum angle and maximum angle values. Each point has x, y and z coordinates in the cloud and the range and angle values for each point are found by simple mathematical operations which are;

$$angle[i] = \arctan(z,x) \qquad \text{and} \qquad range[i] = x^2 + z^2 \qquad i = 1,2,...,N \quad \textbf{(2.9)}$$

where i denotes the index of points, and N denotes the number of points in the cloud

The other important detail is to control the height value of each point by comparing its y position with a prespecified height range. The object in the raw image (Fig. 2.6 (a)) is detected as an obstacle by converting its point cloud (Fig. 2.6 (b)) to the 2D laser scan result (Fig. 2.6 (c)), and the obstacle and its size can be determined by analyzing the result, and then the most available path is identified in order to inform the robot about it.

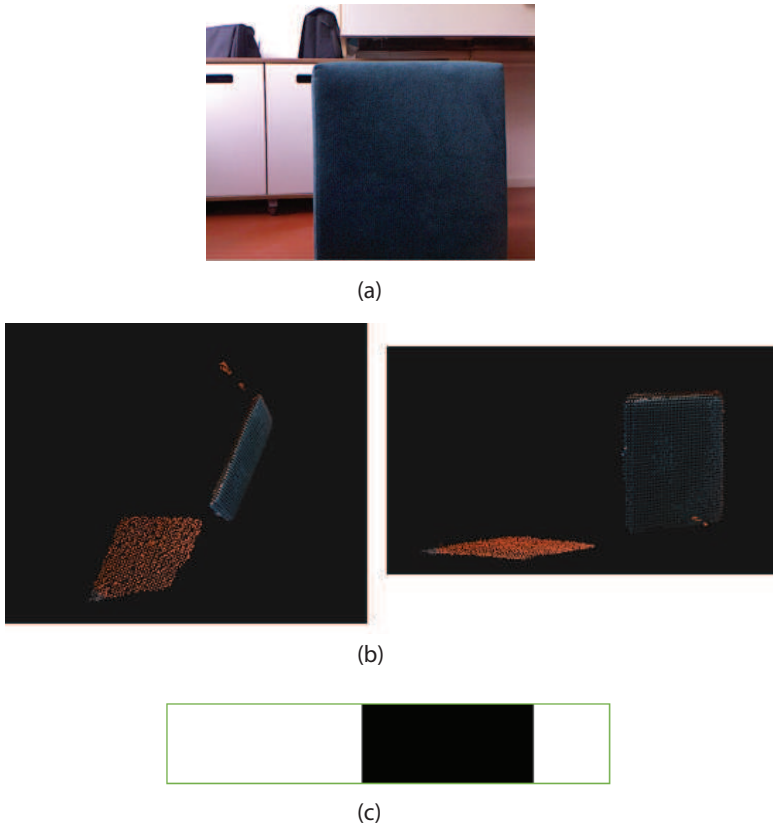

Figure 2.6: A sample conversion from a point cloud to 2D laser scan. In the raw image (a), an object is used to derive its point clouds in (b) to be converted to 2D laser scan result like the pattern shown in (c) where black part representing the obstacle and white parts representing the available floor.

2.1.3 Sensor fusion framework

To reduce uncertainity about sensory data, information obtained from multiple sensors is fused for increasing the quality of the perception. The fusion process leads to obtain more accurate and more reliable results. If the aim of the perception is to track a speaker, auditory and visual data obtained from the person are integrated to improve the continuity in tracking by using probabilistic tracking algorithms such as

particle filtering, Kalman filtering and Extended Kalman filtering. In this research, the particle filtering is utilized to combine the localization outputs from audition and vision modalities to provide more reliable tracking.

2.1.3.1 Particle filter

The particle filtering used for tracking is based on estimating posterior information about the tracked object by using prior information and observations obtained from sensors. A set of particles are trained to find the best posterior information which is only depend on the previous state of the particles and current observation.

Sound source localization from audition modality is integrated with the face localization from vision modality by using a particle filtering based sensor fusion, and the output of sensor fusion is transmitted to the motion modality. Therefore, the robot can prevent losing the track of the source when not obtaining any auditory data. In addition, the displacements of the robot because of obstacle avoidance can be integrated with the output of sensor fusion before obstacle detection if the target is lost due to the displacements.

This probabilistic tracking technique has advantages over Kalman filtering based tracking of an object from a single sensor having no requirement of linear and Gaussian based data [34]; and over Extended Kalman filtering in the cost of computations [35] being proper for non-Gaussian and non-linear auditory data.

When a sound source powerful enough is detected and localized, a number of particles is initialized such as setting the weight of each particle to zero and equating the position of each particle, x, to the localization result, θ. The positions of the particles represent the posterior prediction of speaker position.

$$x_t^k = \theta_t, \qquad k = 1,2,...K \tag{2.10}$$

where t denotes time, and k represents the index of the particles and K is the number of particles.

The positions of initialized particles are sampled by adding a Gaussian distribution-based noise.

$$x_t^k = x_{0:t-1}^k + N(0,\sigma), \tag{2.11}$$

where $N(0,\sigma)$ represents a Gaussian distribution noise having zero-mean and a standard deviation based on the experiments.

The weights, w, of particles representing the proximity to the source localization result are updated according to a Gaussian distribution given observed localization results, θ, and the previous particle position, $(x)_{t-1}$ as below.

$$w_t^k = p(x_t^k | x_{t-1}^k, \theta_t). \tag{2.12}$$

In the proposed multi-modal system, a joint conditional probabilistic distribution is utilized given observed sound source location and visual position of a speaker at time $t-1$.

$$w_t^k = p(x_t^k | x_{t-1}^k, (\theta_a)_t) p(x_t^k | x_{t-1}^k, (\theta_v)_t). \tag{2.13}$$

where θ_v, the face localization output and θ_a, the sound souce localization output are based on the azimuthal angles in the spherical coordinate system at the time, t.

The weights are normalized, so the sum of all weights is 1, and then the resampling step is applied on the normalized weights by sorting them for replacing the smaller half having negligible weights with the larger half having higher proximity to the observations.

Among these particles, the best one is determined according to its weight and its position, θ_{sf}, and is sent to the motion modality. In this work, the localization results for auditory data and visual data are azimuth-based angular value and the angular value on the x-axis in the camera scene, respectively.

If an obstacle is detected while tracking a person and approaching to him/her through the best particle position, the avoidance movements are achieved independently from the output of sensor fusion. After avoiding the obstacle, the last positions of the particles are updated according to the displacements due to angular and linear movements.

$$\begin{aligned} w_{t'}^k &= p(x_{t'}^k | x_{t'-1}^k, \Delta_{t'}^{avoid}) \\ &= \frac{1}{\sigma\sqrt{2\pi}} e^{-\left((x_{t'-1}^k - \Delta_{t'}^{avoid})^2\right) / 2\sigma^2} \end{aligned} \tag{2.14}$$

where t' denotes the time when avoiding from an obstacle and Δ_{avoid} is the displacement value computed during avoiding obstacle, to be utilized to update the particles.

2.2 Active Robot Behaviors

By utilizing the transmitted sensor fusion outputs, robot movements are determined in order to provide more natural interaction with humans. These behaviors are angular movements to detect a speaker and linear movements to approach the visually detected speaker while avoiding from the detected obstacles. The avoidance behaviors and the proximity-based approaching behaviors are not performed at the same time not to affect each other. Thus, the total behaviors of the robot while interacting with a person is expressed as;

$$\begin{aligned} active\_robot\_behaviors = \alpha * approaching\_behaviors \\ + (1-\alpha) * avoidance\_behaviors \end{aligned} \quad (\mathbf{2.15})$$

where α is 0 if the robot performs avoidance behaviors, and 1 if the robot approaches to the tracked person.

2.2.1 Proximity-based approaching behavior

When the behaviors are created, the proximity principle proposed by Newcomb [36] working on social psychology is considered. This theory suggests that humans who are near to each other get in contact more efficiently than people who are further apart from each other. However, most people prefer to keep a particular distance as their personal space and do not like when a human interferes with this area while talking. This dictates practically that it should be mostly avoided to enter the personal area of a human (roughly $\leq$ 50-100 cm in radius), who is contacted by an alien person. By applying this principle to the robotic domain, the robot is supposed to detect, track and approach a speaker while still keeping out of the personal space.

Depth range almost 1 meter [37] is utilized to keep the robot out of the personal space. The robot approaches to the person while not sensing any depth information about the detected face.

2.2.2 Obstacle avoidance behavior

The other intelligent behavior is to avoid the detected obstacles. Firstly, a 2D fake laser scan image which pixels indicate the occupied and free place is found by being converted from a 3D point cloud, and then the ratio of availability on the left and right side are compared. If the difference is bigger than a specified ratio, the motion modality is informed about the more available side to turn the robot to this side until not detecting any obstacles or reaches to a determined angle. If the side does not have enough free space to move the robot linearly, the robot tries to turn to the other side. Each displacement due to the avoidance is used in an update step of sensor fusion as an azimuth position of a person. When the robot finds an available path, it moves linearly for a while, and then it rotates to find the person lost through the best particle position. The displacements are initialized by equating with the position of the best particle, θ_{sf} at that time, and are updated as;

$$\Delta_{t'}^{avoid} = \Delta_{t'-1}^{avoid} - (\alpha * \Delta_{t'-1}^{ang} + (1-\alpha) * \Delta_{t'-1}^{lin}) \tag{2.16}$$

where α is 0 if the robot moves linearly and 1 if the robot performs angular movement, and the value of displacements, Δ^{ang} and Δ^{lin}, of the movements representing the changes in the position if a linear movement or an angular movement is performed are known. The displacements are subtracted because the robot needs to perform only opposite angular movements after avoidance behaviors, in order to find the person.

2.2.3 Selective human based behavior

The proximity based approaching behaviors are performed on a specific person determined through the face recognition, and the observations obtained from another person are ignored when updating the particles in the sensor fusion framework. Therefore, the robot interacts with the recognized person.

3. EXPERIMENTS AND RESULTS

In this chapter, performances of vision modality under different cases are tested and the effects of the size of images and using Haar cascade classifier on the skin color extracted area instead of this proposed face detection method are shown in the first experiment. The results of sensor fusion framework are evaluated for a single person speaking while moving and two people speaking simultaneously and independently while not moving. In the last experiment, a real robot is utilized with the multimodal system with a single person and two people in the environment having no obstacles and with a single person in the environment with obstacles.

3.1 Software and Hardware Specifications

An open source library called Open Computer Vision Library[1] (OpenCV) created and developed by Intel to simplify the usage of requirements in computer vision is utilized in the vision module. The facilities provided by the library, used for preprocessing are morphological operations, gray-scaled conversion, and Histogram Equalization. The other facilities required in this project are Haar Cascade Classifier, Camshift algorithm, recognition based on eigenfaces, drawing shapes on images like rectangles and circles, putting texts on the images, the operations for resizing and cropping the image frames. We used the following data types which are **Mat** data type to represent images, **Rect** data type for detected faces and eyes, **Point** data type to be used for specifying the boundaries of required rectangles, circles and position information of detected faces.

To process depth information to predict the distance of the robot to a speaker and to convert point clouds to a virtual 2D laser scan, Point Cloud Library[2] (PCL) is utilized.

For the audition modality, sound source detection and localization processes are achieved by using HARK[3] (HRI-JP Audition for Robots with Kyoto University)

[1] http://opencv.org/
[2] http://pointclouds.org/
[3] http://www.hark.jp/

audition software for robotics. This software provides modules implementing techniques to be used in signal processing. Networks are created for sound recording, sound source localization, sound source separation, speech recognition and relevant operations.

Robot Operating System[4] (ROS) is used as a common framework for both modalities in order to assemble the localization results for fusion and then to transmit to the motion modality. ROS provides open-source packages about robots, sensors such as Kinect, libraries like OpenCV and Point Cloud Library, simulators and algorithms used in robotics, special data types, commands for command-line interpreter and Subscribe/Publisher facility for signalling between nodes by using rostopic buffers. The data flow between the nodes of modalities in the system through the rostopics determined in the packages or created by users is achieved.

In this project, C/C++ programming language is utilized within these libraries, HARK modules and the ROS framework.

Hardware used in this project are two Kinect (Fig 3.1) sensors. One is required for obtaining video streams to be used in face detection and the other is equipped on the bottom of the cast of Turtlebot to be used to obtain point clouds for obstacle detection.

Figure 3.1: Kinect camera sensor.

Microcone microphone array (Fig 3.2) having 7 channels for capturing audio streams and a modified version of Turtlebot II (Fig 3.3) with almost the height of a child (150 cm) as a mobile robot are utilized. The multimodal system is run on two netbooks each with 4 GB RAM and 1.5 GHz Intel Celeron dual core CPU. One is for data processing in the system, and the other for communication with Turtlebot and data processing for obstacle detection.

[4]http://www.ros.org/

Figure 3.2: Microcone microphone array.

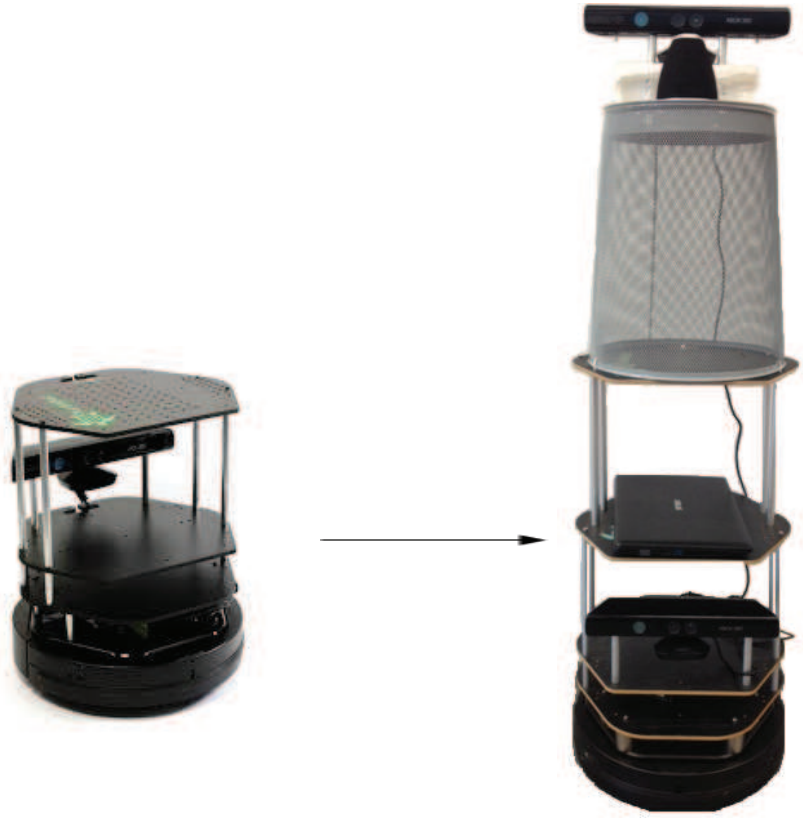

Figure 3.3: Original Turtlebot version II and its modification for this research.

3.2 Experimental Setup

Auditory/visual recording and real-world experiments are carried out in Istanbul Technical University HP Laboratory. Human localization using visual cues is performed on a video stream having 10 frames-per-second (fps). The recordings conducted on the vision modality includes a person turning his face to different positions -sometimes looking and sometimes not looking to the camera directly- while simultaneously moving in the room. In this experiment, the performance of the human

tracking using vision modality is tested and detection accuracy of each individual step of the proposed method is evaluated (Section 3.3.1).

The dataflow of the audio processing system runs at time increments of 10 ms, using a Complex window of 512 samples and 32% overlap (*i.e.*, hop size of 160 samples) for computing the audio spectrum. The performance of auditory and audio-visual human tracking are tested in Section 3.3.2 with two recordings, one having only one person speaking while moving and one having two people speaking together as well as independently. Also, the vision modality performance with multiple persons is tested. In the experiments, total number of particles, K, is selected as 200 and standard deviation, σ, is selected as 15 for sensor fusion.

In the last experiment in Section 3.3.4, robot behaviors in the framework of active robot perception are assessed. If the speaker is not in the robot eye-sight, the robot follows motion commands of *angular rotation* by the contribution of sound source localization results, which dictate the robot to search for a face. When a face is detected, initialized particle positions are updated and motion commands for *linear movement* are sent to the robot according to the distance to the detected face.

The tracking and approaching behaviors are based on the position of the best particle transmitted from the fusion framework to be used to center the speaker. Before implementing sensor fusion, the x value of the center point of each detected face which is a pixel position on the image is converted to an azimuthal angle in the spherical coordinate system in order to integrate with a sound source localization output, θ_a which is also an azimuthal angle, and to update the particle positions when only visual data is observed. The microphone array captures sounds from the angles between -180° and 180°, and so the angles in the eyesight of the camera are between -60° and 60° while the person is far away enough from the camera where his/her speech can be detected. The x value is converted to obtain its angle version, θ_v as;

$$\theta_v = -\frac{c_x * 120}{W} + 60 \qquad \textbf{(3.1)}$$

where c_x denotes the x value of the center point which is between 0 and W which is the width of the image as 320.

However, the angles of the microphone array between -60° and -180°, and between 60° and 180° are not in this eyesight, and so if the robot detects and localizes a sound source in the angles, it needs to rotate to search a face in this location as shown in the last experiment.

3.3 Results

In this section, the results of each step in the proposed face detection method in different cases and records are shown and the contribution of the sensor fusion framework into the sound source localization performance is evaluated.

3.3.1 Results of face detection experiments

In the first experiment, the vision modality is tested using 3 recordings lasting 2 minutes and having different lighting conditions (Fig. 3.4), different background colors, and a person changing the orientation of his face while moving. Precision and recall values are compared between 1) Haar Cascade Classifier-based face detection without eyes detection, 2) a face detection with at least one eye detection without skin-color detection, 3) a face, eyes and skin-color detection and 4) the proposed method including Camshift.

In the experiments, the details are also considered to improve the performance of face detection in real-time, which are to select the scale factors on Haar-feature detection for eyes and face separately and to determine the size of a frame. In table 3.1, the face detection results are shown in three different size pairs at the same fps by adjusting the scale factors. Therefore, the frame size where width and height of the size are 320 pixels and 240 pixels, respectively, the size of searching area for eyes and face detection is selected as 1 pixel for width and 1 pixel for height, and scale factors, 1.05 for searching face features and 1.01 for searching eye features are selected by considering time complexity.

However, face recognition is also based on the size of a face, therefore width and height values are selected as 320 and 240 respectively, and search factors for the face and the

WidthxHeight	only face	face + eyes
1280x960	39	14
640x480	269	118
320x240	638	434

Table 3.1: The number of the faces detected with different sizes of the same frame and the scale factors selected not to affect fps for each frame size, which are 2.4, 1.3 and 1.05 for face features, 1.8, 1.15 and 1.01 for eye features respectively, are compared to find appropriate frame size with scale factors.

eyes in a frame are selected as the smallest one in order to detect faces far away from the camera.

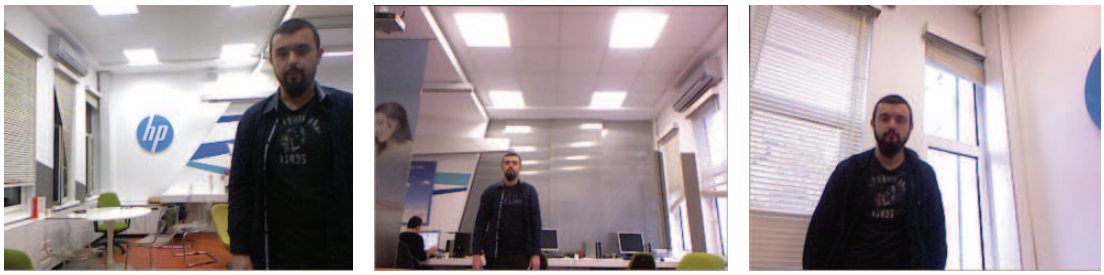

Figure 3.4: Three recordings used to evaluate the performance of the proposed face detection method in vision modality.

Table 3.2: Face detection results.

Processing Method	Results			
	False Positive	True Positive	Precision	Recall
Only Face	340	226	0.39	0.20
Face and Eye	65	199	0.75	0.18
Face, Eye and Skin-color	0	199	1.00	0.18
Proposed Method	12	1023	0.98	0.94

Table 3.2 shows the number of false positive detection results, true positive detection results, precision and recall values on the three recordings having altogether 1100 frames. By using only face detection based on Haar features, false positives are found to be 340 times, and by searching eyes on that false positives, most of them are eliminated.

With the additional skin-color constraint, all false-positives are eliminated and it is observed that the entire true positives have skin-color. However, the face is detected only 199 times out of the 1100 frames. It is not sufficient for tracking, even it has the perfect precision value. Therefore, when a face with eyes and having skin-color

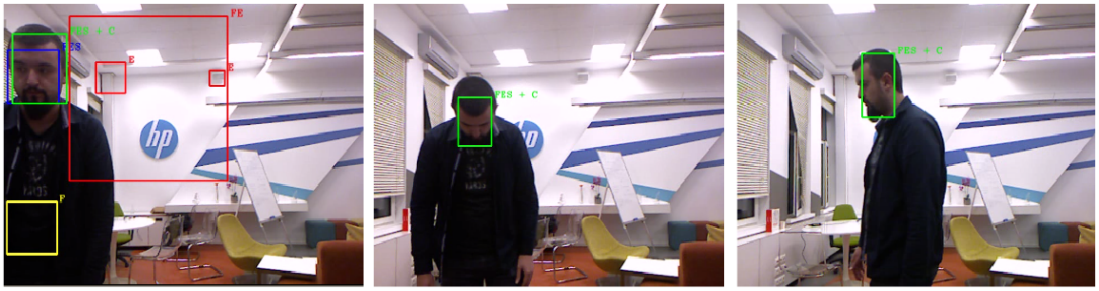

Figure 3.5: Face detection and tracking results. F: Face detection, E: Eye detection, S: Skin-color constraint and C: Camshift algorithm.

is detected, its position is sent to Camshift in order not to lose the face in an adaptive YCbCr color space. In this method, the instances in which Camshift algorithm returns abnormal rectangles because of unexpected changes in lights are counted as false positives. Some true positives are missed because of pausing the Camshift algorithm.

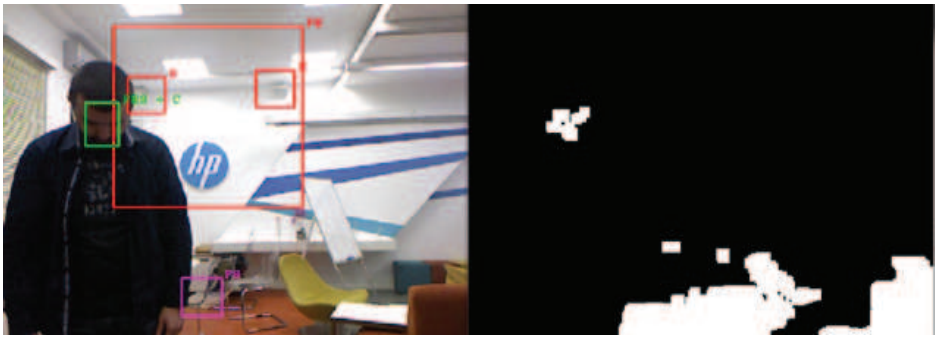

Figure 3.6: Using Haar-feature based eye detection eliminates some false-positives like the pink rectangle representing facial area having skin-color.

In Fig 3.5; false positive samples of face detection by Haar cascade classifier with and without eye detection are shown. The proposed method clearly demonstrates a better accuracy of face detection than the other steps. Moreover, even though the face poses are not frontal, which is definitely affecting the Haar-feature detection, Camshift algorithm still yields correct results if the skin color is extracted correctly as shown in this figure. In Fig 3.6, the advantage of using eye detection is illustrated. It eliminates false positives of Haar-feature based facial areas having skin-color such as in the pink rectangle. However, 27 true-positive samples are missed because of the distance to the camera or lighting conditions.

A common method used for face detection is that a background subtraction is applied by extracting skin areas and then feature based detection techniques are implemented

on the areas (Fig 3.7). By using this process, the accuracy of eye detection is decreased because the area including eyes does not have skin color. Therefore, the number of true positives is smaller and the number of face detections are more (Table 3.3). However, most of the detected faces is false positive, and so the importance of eye detection is evaluated in eliminating false positives.

	face + eye	**face + eye + skin**	**face + skin**
Skin Extraction + Feature	58	56	173
Feature + Skin Extraction	537	109	129

Table 3.3: The number of the detected faces by using feature based detection and then skin color constraint, and skin color extraction and then feature based detection.

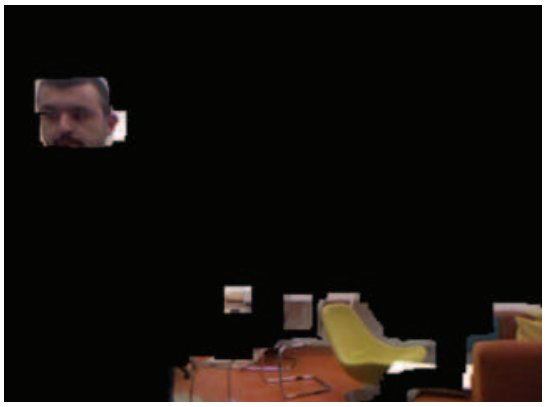

Figure 3.7: Haar-features are searched on Skin color extracted frames.

3.3.2 Results of audio-visual human tracking experiments

The performances of audition modality and sensor fusion are examined on two recordings; one having only one person talking while moving and the other including two people talking. Sound source localization is required for robots especially in order to start the interaction with humans and to make them turn towards the direction of the potential location of a talker when there is no human in the eye-sight of robots.

In Fig. 3.8, it is shown that the person moves from the right (-30°) side of the robot to the left side (30°) while talking. In this figure, sensor fusion decreases an error of sound source localization of 10° to almost 5°. Around 30.3 seconds of this 45 seconds-long

recording[5] has sound activity. The average error of sound source localization is 6.13° and the sensor fusion decreases this error to 1.86° in average.

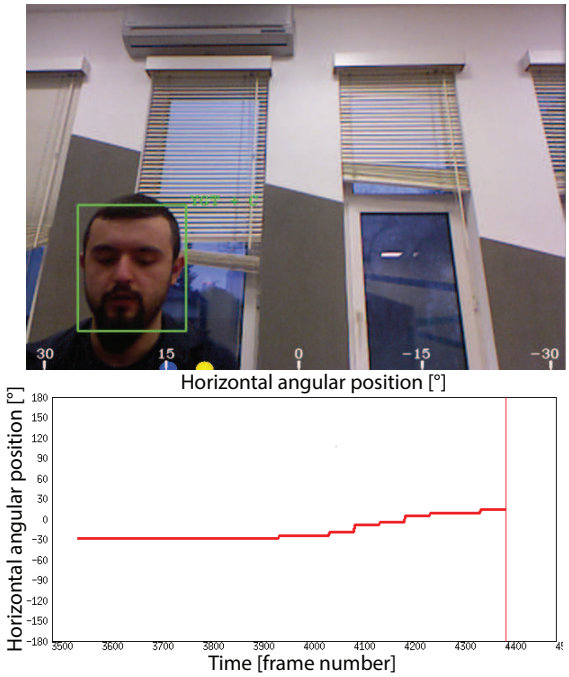

Figure 3.8: Audio-visual human tracking implementation. In the panel above, the yellow dot represents angular position of sound source and blue dot represents the estimated human position from sensor fusion. The panel below shows the sound source localization results over time.

In another recording with two persons (Fig. 3.9), it is shown that two persons speak at the same time on the right (-40°) side of the robot and the left side (40°). In this figure, sensor fusion decreases 7.5° error of sound source localization to almost 1° for the left speaker, and sensor fusion decreases 10° of sound source localization error to 3° for the right speaker.

The entire 45 seconds recording[6] including two persons speaking separately at first, and then speaking together for a while includes sound activities. The average error of sound source localization is 3.77° and the sensor fusion decreases this error to 1.4°. In addition, the proposed method in the vision modality for visual tracking is observed when multiple people are in the eye-sight. Because of the proximity and view of face in the camera angle, the method performs without any false positive for each face.

[5] http://youtu.be/oXZ88vAYM6o
[6] http://youtu.be/21qCSpeWxgk

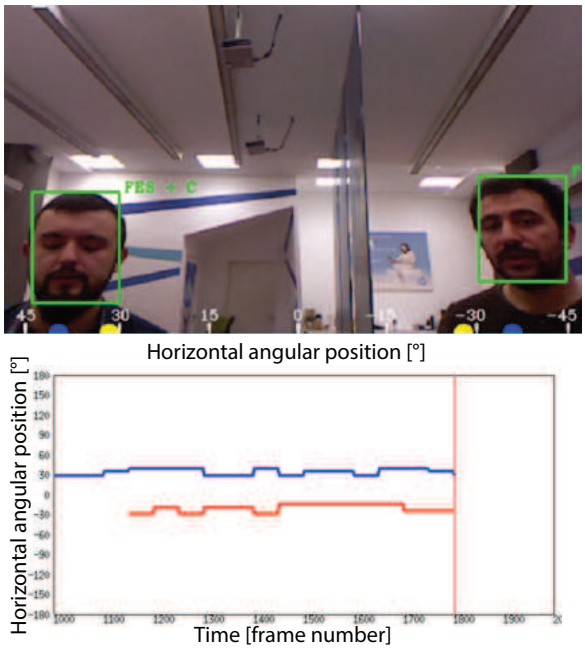

Figure 3.9: Audio-visual human tracking implementation. In the panel above, the yellow dots represent angular positions of sound sources and blue dots represent the human positions from sensor fusion. The panel below shows the sound source localization results over time.

3.3.3 Results of audio-visual selective human tracking experiments

In this experiment (Fig. 3.10, the performance of the sensor fusion is observed. The fusion uses only the observations obtained from the recognized person on the right side, and the auditory and visual observations of the other person are ignored.

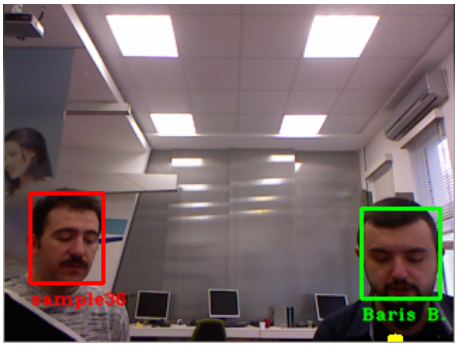

Figure 3.10: Audio-visual selective human tracking implementation. The yellow dot shows the sensor fusion output.

3.3.4 Results of active robot perception based behaviors experiments

In the last experiment, the robot behaviors are shown in three real-world experiments (Fig. 3.11[7]) and (Fig. 3.12[8]). It is observed that the robot tracks a speaker changing his face poses while moving and navigates between two speakers while taking the personal distances into consideration. In the experiment with multiple person, audition modality is mostly required and robot behaviors angularly moving to each detected sound source are shown, and the vision modality is mostly utilized to track the single person while moving.

Figure 3.11: Audio-visual human tracking applied to a person while moving in a real environment. From (a) to (b), the speaker moved his chair in the room.

In the Fig. 3.13 for the third experiment, proximity-based behaviors and obstacle avoidance behaviors are performed separately in this experiment. The robot detects an obstacle while approaching to the person and rotates to the available side to avoid

[7]http://youtu.be/QOWi66kJkTI, http://youtu.be/vhvGvDJ2wYQ
[8]http://youtu.be/JkJeNhm7jsY

Figure 3.12: Audio-visual human tracking applied to multiple people in a real environment.

from it and then performs opposite angular movement through sensor fusion output in order to find the person to complete its approaching behavior.

Figure 3.13: Audio-visual human tracking applied to a person in a real environment with an obstacle.

The most common problems affecting robot motions in real experiments are observed such as the following: 1) noise sources were detected and localized instead of useful sound sources, 2) Camshift algorithm focused to areas similar to skin-color for a while, and 3) the transmissions of the motion commands to the robot were sometimes missed because of the high dataflow and processing frequencies.

To overcome these problems, we increased predefined threshold value due to the noise sources, used a lifetime for Camshift until a true positive is detected, and sent a few more motion commands to the robot.

4. CONCLUSION

In this thesis, auditory and visual perception for active robots to detect and localize humans, and intelligent behaviors due to the perception results to provide natural interaction are investigated and implemented with a real robot in real environments with obstacles.

Sound source localization is an essential ability for robot audition, but the method has problems such that the existance of noisy sounds more dominant than the target sound source affects the location of the detected sound sources. Audition modality in our study provides sound source detection and localization by using MUltiple SIgnal Classification based on GEVD-MUSIC. This method can find the powerful sound sources in the environment by eliminating noisy sounds, in case the noise can be predicted in advance.

The method proposed in the vision modality covers four steps by using Haar cascade classifier for face and eyes and YCbCr color space used for reliable face detection to ensure that the detected area is a face indeed. These steps are 1) face detection, 2) (at least one) eye detection on the face area and 3) a constraint check on the skin-color proportion under YCbCr color space on the detected face with eyes area. If this detected area has sufficient amount of skin-color, the coordinates of the area are further sent to 4) Camshift the color-based tracking algorithm to track the detected face. Before sending the detected facial area to Camshift, eigenface based face recognition is applied to determine the identity.

The results from both modalities are integrated by the sensor fusion framework to be relayed to the motion modality. This framework consists of a particle filter-based position estimation technique to integrate the results from sensor modalities. Active robot behaviors are proposed such as tracking a speaker and approach of the robot to the person by taking the proximity principle into account. By using the motion modality, a robot may be equipped with intelligent behaviors like tracking a speaker

and approaching to the person only up to a distance not exceeding the optimal radius of a human's personal space in order to avoid disturbing the user.

When the robot moves to approach the detected person, obstacle detection is achieved to avoid the obstacles by using a virtual 2D laser scan from point clouds and the sensor fusion output is updated considering the avoidance movements to find the person after finding available path.

As a future work, we intend to implement additional abilities to the robot like integrating auditory and visual data to achieve intelligent automatic speech recognition and to design new behavior patterns considering the localization results of multi-speaker talking at the same time to be used in sound source separation. Likewise, another navigational tasks are considered such as mapping, and it is purposed to achieve another robot behaviors such as robotic arm movements with respect to the audio visual integration outputs.

REFERENCES

[1] **Nakadai, K.**, 2006. Robust Tracking of Multiple Sound Sources by Spatial Integration of Room and Robot Microphone Arrays, *Proc. of IEEE ICASSP*, **IV**, 929–932.

[2] **Nakadai, K.**, **Ince, G.**, **Nakamura, K. and Nakajima, H.**, 2012. Robot Audition for Dynamic Environments, Proceedings of the IEEE International Conference on Signal Processing, Communications and Computing (ICSPCC), pp.125–130.

[3] **Hara, I.**, 2004. Robust Speech Interface based on Audio and Video Information Fusion for Humanoid HRP-2, Proceedings of IEEE/RSJ International Conference on IROS, pp.2404–2410.

[4] **Kim, H.**, 2007. Auditory and Visual Integration based Localization and Tracking of Humans in Daily-life Environments, Proceedings of the 2007 IEEE/RSJ International Conference on IROS, pp.2021–2027.

[5] **Murray, J.C.**, **Erwin, H. and Wermter, S.**, 2004. Robotic Sound-Source Localization and Tracking Using Interaural Time Difference and Cross-Correlation, Proceedings of NeuroBotics Workshop.

[6] **Viola, P. and Jones, M.** Rapid object detection using boosted cascade of simple features, IEEE Conference on Computer Vision and Pattern Recognition, pp.511–518.

[7] **Shen, B.C.**, **Chenl, C.S. and Hsu, H.H.**, 2008. Face image retrieval by using Haar features, 19th International Conference on Pattern Recognition (ICPR), pp.1–4.

[8] **Wilson, P. and Fernandez, J.**, 2006. Facial Feature Detection Using Haar classifiers, *JCSC*, 127–133.

[9] **Singh, S.K.**, **Chauhan, D.**, **Vatsa, M. and Singh, R.**, 2003. A Robust Skin Color Based Face Detection Algorithm, *Tamkang Journal of Science and Engineering*, **6(4)**, 227–234.

[10] **Niazi, M. and Jafar, S.**, 2010. Hybrid Face Detection with HSV Color Method and HAAR Classifier, International Conference on Software Technology and Engineering - ICSTE.

[11] **Maghraby, A.E.**, **Abdalla, M.**, **Enany, O. and El Nahas, M.Y.**, 2013. Hybrid Face Detection System using Combination of Viola - Jones Method and

Skin Detection, *International Journal of Computer Applications*, **71(6)**, 15–22.

[12] **Wang, J. and Tan, T.**, 2000. A new face detection method based on shape information, *Journal of Pattern Recognition Letters*, **21(6-7)**, 463 – 471.

[13] **Yuille, A.L.**, **Hallinan, P.W. and Cohen, D.S.**, 1992. Feature extraction from faces using deformable templates, *International Journal of Computer Vision*, **8(2)**, 99 – 111.

[14] **Wren, C.**, **Azarbayejani, A.**, **Darrell, T. and Pentland, A.** Pfinder: Real-Time Tracking of the Human Body, IEEE Transactions on Pattern Analysis and Machine Intelligence, pp.780 – 785.

[15] **Huang, J.**, **Gutta, S. and Wechsler, H.** Detection of Human Faces Using Decision Trees, International Conference on Automatic Face and Gesture Recognition, pp.248 – 252.

[16] **Cheng, Y.**, August 1995. Mean Shift, Mode Seeking, and Clustering, *IEEE Transactions on Pattern Analysis and Machine Intelligence (IEEE)*, **17(8)**, 790–799.

[17] **Bradski, G.R.**, 1998. Real Time Face and Object Tracking as a Component of a Perceptual User Interface, WACV '98 Proceedings of the 4th IEEE Workshop on Applications of Computer Vision (WACV'98), p.214.

[18] **Jacquin, A. and Eleftheriadis, A.** Automatic Location Tracking of Faces and Facial Features in Video Sequences, International Workshop on Automatic Face and Gesture Recognition, pp.142 – 147.

[19] **Hua, R.**, **De Silva, L. and Vadakkepat, P.**, 2002. Detection and Tracking of Faces in Real-Time Environments, Proceedings of International Conference on Imaging Science, Systems and Technology (CISST '02).

[20] **Turk, M. and Pentland, A.**, 1991. Eigenfaces for Recognition, *J. Cognitive Neuroscience*, **3(1)**, 71–86.

[21] **Umbaugh, S.**, 1998. Computer Vision and Image Processing, Prentice Hall, NJ.

[22] **Gonzales, R. and Woods, R.**, 2002. Digital Image Processing, Prentice Hall.

[23] **Yoshida, T.**, 2009. Automatic speech recognition improved by twolayered audio-visual integration for robot audition, *Proc. of Humanoids*, 604–609.

[24] **Koiwa, T.**, 2007. Coarse speech recognition by audio-visual integration based on missing feature theory, Proceedings of the 2007 IEEE/RSJ International Conference on IROS, pp.1751–1756.

[25] **Nakamura, K.**, **Nakadai, K.**, **Asano, F. and Ince, G.**, 2011. Intelligent Sound Source Localization and Its Application to Multimodal Human Tracking, Intelligent Robots and Systems (IROS), pp.143–148.

[26] **Pavlovic, V. and Huang, T.S.** Multimodal Tracking and Classification of Audio-Visual Features, AAAI Workshop on Representations for Multi-modal Human-Computer Interaction, pp.343 – 347.

[27] **Nakadai, K.**, 2000. Active audition for humanoid, Proceedings of National Conference on Artificial Intelligence (AAAI), pp.832–839.

[28] **Rodemann, T.**, **Ince, G.**, **Joublin, F. and Goerick, C.**, 2008. Using binaural and spectral cues for azimuth and elevation localization, Intelligent Robots and Systems (IROS), pp.2185–2190.

[29] **Berglund, E. and Sitte, J.**, 2005. Sound source localisation through active audition, Proceedings of IROS, pp.653–658.

[30] **Sasaki, Y.**, 2006. Multiple sound source mapping for a mobile robot by self-motion triangulation, Proceedings of IROS, pp.380–385.

[31] **Martinson, E. and Brock, D.**, 2007. Improving human-robot interaction through adaptation to the auditory scene, in ACM/IEEE Int. Conf. on Human-Robot Interaction (HRI), pp.113–120.

[32] **Yamamoto, S.**, **Nakadai, K.**, **Nakano, M.**, **Tsujino, H.**, **Valin, J.M.**, **Komatani, K.**, **Ogata, T. and Okuno, H.G.** Design and Implementation of a Robot Audition System for Automatic Speech Recognition of Simultaneous Speech, IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), pp.111 – 116.

[33] **Salhi, A. and Jammouss, A.**, 2012. Object tracking system using Camshift, Meanshift and Kalman filter, *World Academy of Science, Engineering and Technology*, **6(4)**, 598 – 603.

[34] **Checka, N.**, **Wilson, K.**, **Siracusa, M. and Darrell, T.**, 2004. Multiple person and speaker activity tracking with a particle filter, In International Conference on Acoustics, Speech and Signal Processing. IEEE, pp.881–884.

[35] **Rigatos, G.G.**, 2007. Extended Kalman and particle filtering for sensor fusion in mobile robot localization, PhysCon' 07, IEEE International Conference on Physics and Control, p.328.

[36] **Newcomb, T.**, 1960. Varieties of interpersonal attraction, In D. Cartwright & A. Zander (Eds.), 2. edition.

[37] **Cruz, L.**, **Lucio, D. and Velho, L.**, 2012. Kinect and rgbd images: Challenges and applications, Graphics, Patterns and Images Tutorials (SIBGRAPI-T) in 25th SIBGRAPI Conference Conference, pp.36–49.

APPENDICES

APPENDIX A : The list of ROS commands used in this project
APPENDIX B : The SSL network in HARK and the description of all modules
APPENDIX C : The list of rostopics used in the project

APPENDIX A

The commands in ROS used during the project are;

roscore – to create a master node

rosmake – to compile a source code.to create a node.

roslaunch <package name> <.launch file>– to operate cameras and the robot in order to create relevant rostopics to obtain data from cameras and to transmit data to the robot.

rosrun <package name> <executable file name>– to run the executable files created for each node.

rostopic list – to list all rostopics available in the existing master node.

rostopic echo <rostopic name> - to observe the data flowing in a specific rostopic

rosbag record <rostopic names> - to record data in specified rostopics as a .bag file. It was used to take some visual and sound records to test the proposed system

rosbag play <.bag file> – by using this command, a .bag file is played and then all rostopics in the file are started to flow data.

rxgraph – to create a kind of data flow graph showing the relation between nodes and the rostopics used on these relations.

APPENDIX B

The sound source localization network in HARK shown in Fig B.1 includes the modules required to detect and localize sound sources and publishing module to transmit the SSL results to the sensor fusion node in ROS.

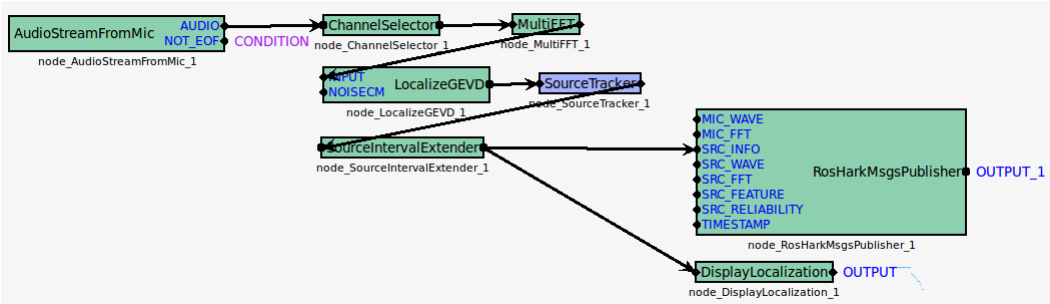

Figure B.1: SSL HARK network used in these experiments

The modules used in this network are;

AudioStreamFromMic obtains multichannel audio data from a microphone array with predefined attributes: channel number, device type, the sampling frequency rate of the data and device.

ChannelSelector is required to select several channels from audio data coming from multiple channels.

MultiFFT is used to perform FFT (Fast Fourier Transform) on the multichannel audio data in order to convert it to spectra in frequency domain.

LocalizeGEVD performs the GEVD-MUSIC localization method.

SourceTracker is used to give the same IDs to the localized sound sources on adjacent direction or different IDs to the sources coming from different directions.

SourceIntervalExtender is required to output source localization results earlier.

RosHarkMsgsPublisher publishes the localization results to ROS with a rostopic to be subscribed by a node.

DisplayLocalization displays the localization result including time and angular position axes.

APPENDIX C

The data flows by using rostopics in the mutlimodal system and relations between nodes implemented in ROS is shown as a graph in Fig. C.1.

The rostopics created are;

/initialized – the data about sensor fusion initialization is sent to the vision module. After this initialization, the face position values are sent to motion module for fusing them with the other observation values.

/face_identity – contains the name from the face recognition process implemented for each true positive detection. This rostopic does not affect the motion of the robot, but it enables the robot to know whom it is tracking.

/proximity – sends the value calculated from depth information to determine the amount of linear robot motions. If it is higher than a predetermined value, which is 0.5, it is too close, and then the backward motion command is sent to the robot, and vice versa.

/face_position – transfers the point information about the face to the motion module to be used in sensor fusion to specify the angular velocity values of the robot.

/sound_source_position – transfers the sound source localization result which is an angle value between -180° and +180° to be converted to the same data type with the face position in order to be fused.

/obstacle_detected – the values from this rostopic is important to prevent the motion commands from two different nodes. If it is true, the linear or angular motion commands are stopped from sensor fusion until the detected obstacle is completely avoided. The commands from the sensor fusion are applied when there are no more any detected obstacles.

/available_path – includes the information about the free side on floor to inform the motion module

The rostopics from sensors are the two rostopics, */rgb/image_raw* and */depth_registered/image_raw* published from Kinect used for face detection and proximity estimation created by using the *openni_node.launch* launch file in the *openni_camera_deprecated* package

/depth_registered/image_raw – The depth information coming from this rostopic is utilized to understand the proximity of the detected face.

/rgb/image_raw – includes raw image information. The proposed face detection system is applied on these information. The bottom Kinect is activated by using *openni.launch* launch file in the *openni_launch* package to create */camera/depth/image_raw* and */camera/depth/points*. The reason of using two different kind of launch file is to create rostopics having different names because ROS

does not permit the same name rostopic creation even they are coming from different computers.

$/camera/depth/image\_raw$ – includes information required to detect an obstacle as a virtual laser scanner.

$/camera/depth/points$ – to detect the obstacle by converting the point cloud data coming with 1ms from this rostopic to a fake laser scan.

$/HarkSource$ – is created by RosHarkMsgsPublisher module in SSL HARK network to send the information about sound source localization to sensor fusion node, which is azimuth angle

$/cmd\_vel\_mux/input/teleop$ – created by using the $minimal.launch$ file in $turtlebot\_bringup$ package to make the mobile robot Turtlebot available to move. This rostopic includes linear velocity and angular velocity values coming from motion module in order to transmit the robot. Therefore, by using this facility, how many times the robot motion commands sent to the robot by using this rostopic are specified instead of distances. Because of this limitation, the perception results and motion commands are conflicted and some commands on the rostopic are missed by the robot.

The angular velocities transmitted with rostopic $/cmd\_vel\_mux/input/teleop$ to Turtlebot are computed according to the position values from only audition modality. These values should be bigger than 0.6 because it is the minimum value achieved by the robot without missing a movement. Moreover, the values should be less than 2.0 not to cause physical problems in the robot. Therefore, the values are compared with these threshold values and if they are not in the range, they are normalized to be equal to the thresholds.

$$x_t^k = 180 \times \theta / 2\pi \tag{A.1}$$

By using sensor fusion results, the signs of angular velocities are determined through the distance to the mid-point on the eyesight of the robot and their values are computed by using the distance to the mid-point.

For linear movements, the linear velocity is determined as 0.1 to move smoothly and troubleless.

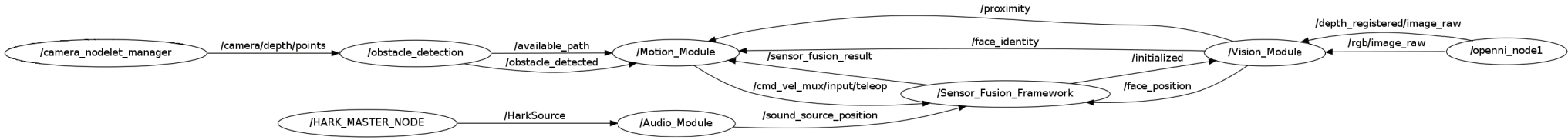

Figure C.1: The multimodal system implemented in ROS

CURRICULUM VITAE

Name Surname:
Barış Bayram

Place and Date of Birth:
25.06.1989

Adress:
Kaptan Sok. 27/4 Yılmaz Apt. Çeliktepe İSTANBUL

E-Mail:
brsbyrm@gmail.com

B.Sc.:
Izmir University of Economics, 10.2007- 06.2012

M.Sc.:
Istanbul Technical University, 09.2013 – ...

Professional Experience and Rewards:
TRR Robot ve Bilişim Teknolojileri, 08.2014-01.2015 as R&D Engineer

PUBLICATIONS/PRESENTATIONS ON THE THESIS
B. Bayram, G. Ince: Aktif Robot Algılaması İçin Görsel-İşitsel İnsan Takibi, Proceedings of the 23rd Signal Processing and Communications Applications Conference (SIU 2015), 2015