

An Application of Passive Human–Robot Interaction: Human Tracking Based on Attention Distraction

Ali Şafak Sekmen, *Member, IEEE*, Mitch Wilkes, *Member, IEEE*, and Kazuhiko Kawamura, *Fellow, IEEE*

Abstract—There is much interest in developing methods that enable humans and robots to interact in a natural and unencumbered fashion. In natural human–robot interactions, the human can be considered as a passive user since he/she does not have to behave in an artificial manner. The robot is in the direct physical presence of the human, who does not need to use any unnatural interfaces such as the keyboard or mouse to interact with the robot. In this research, a taxonomy is introduced to cover important considerations for human–robot interactions. As an application of passive human–robot interaction, two modalities for localizing humans based on sound source localization and infrared motion detection have been developed and integrated with the face-tracker system of a humanoid, Intelligent Soft Arm Control (ISAC), in order to direct ISACs attention and to prevent it from being quickly distracted. The sound source localization and passive infrared (PIR) motion detection systems are used to provide the face-tracker system with candidate regions for finding a face. In order to avoid the situation where the robot appears to be “hyperactive” and cannot give sufficient attention to a newly discovered face, these sensing modules should not directly gain control of the tracking if the system has recently acquired a new face. Our goal is to allow a human to redirect the attention of the system but give the system a method to ignore the distraction if recently engaged.

Index Terms—Face-tracker, infrared motion detection, passive human–robot interaction, service robots, sound source localization.

I. INTRODUCTION

ALTHOUGH robots seem to possess excellent skills in science fiction movies, it would be surprising to many people to learn how many improvements today’s service robots need in order to be capable of doing relatively simple tasks. The interaction between two people is natural since they understand each other, in contrast to the interaction between a robot and a human. The focus of this research is on the development of new passive human–robot interaction techniques to improve the current interactions systems of the humanoid system, ISAC.

If a robot is in a direct contact with a human, it is desirable for it to have some natural communication means such as voice or gesture recognition systems. The robot should imitate a human in some ways. For example, it should have ears to localize a person speaking around the robot, and it should even be able to understand what the person is saying. It should also have a

vision system for looking around and extracting features as a human does. In addition, the robot should be able to track the person as he/she moves around and talks. The robot should not be restricted to human-like capabilities and it may have some sensing mechanisms that humans do not. For example, it is not possible for a human with closed eyes to detect and track a moving person when that person does not make any noise. However, a robot equipped with an infrared sensor array can sense human movement and localize him/her.

In direct interaction, the robot and the user are in the same environment (i.e., immediate location). Here, the human can be considered as a passive user since he/she does not have to behave in an artificial manner, but naturally. When a human enters a room in which the service robot is placed, it is enough if he talks to the robot as if he is talking to another person. For example, he can say “hello” and the robot can localize the user by its sound source localization system and look at him and reply back by saying “hello.” The user does not even need to talk. When he enters the room, the robot can detect his movements and track him by its cameras using its passive infrared motion detection system. Therefore, in the remainder of this paper, direct interaction shall henceforth be referred to as passive interaction.

It is very important for a service or mobile robot to be able to localize or track people it is serving or working around. For example, a service robot first needs to localize the person to whom it is supposed to give a cup of water. The localization process can be done through different technologies such as video, infrared subject tracking or sound source localization. Each of those technologies has its own advantages and disadvantages.

Using audition instead of vision has some advantages. First of all, it does not require any illumination and it is practically omnidirectional. Second, a sound signal is a one dimensional signal requiring significantly fewer computational resources compared to image-based systems [1], [2]. Third, audio signals can travel through or around obstacles and give more cues for localization.

The importance of sound localizing devices for these various applications has produced several research efforts on sound source localization. Even though most of the researchers have tried to imitate the human binaural system, they have used different sensor configurations consisting of different numbers of microphones. Brandstein and Wang implemented a face tracking system based on both sound and visual cues. The initial location of the talker is estimated by a sound localizing system composed of four microphones, and then a face tracking algorithm depending on face motion detection is used to track the talker [3]. Sturim *et al.* produced an acoustic microphone array to track multiple talkers [4]. They developed a method for tracking the positional estimates of

Manuscript received April 17, 2001; revised February 28, 2002. This paper was recommended by Associate Editor A. Ollero.

A. S. Sekmen is with the Department of Computer Science, Tennessee State University, Nashville, TN 37209 USA (e-mail: asekmen@tstate.edu).

M. Wilkes and K. Kawamura are with the Center for Intelligent Systems (CIS), Electrical Engineering and Computer Science Department, Vanderbilt University, Nashville, TN 37235 USA (e-mail: wilkes@vuse.vanderbilt.edu).

Publisher Item Identifier S 1083-4427(02)06005-8.

multiple talkers in the operating region of their microphone array. The microphone array is used to estimate initial talker locations by a time-delay based localization algorithm and extended Kalman filtering is used to smooth the estimates. Wang and Chu implemented an automatic camera pointing system using a configuration consisting of four microphones [5]. Their system calculates three-dimensional (3-D) position of a talker: azimuth, elevation, and range, by estimating time delays of four pairs of microphones. Guentchev and Weng used a learning-based approach for sound localization [6]. They use a two-level procedure. In the training part, sound samples with known coordinates are produced and stored. In the recognition part, unknown sound samples are processed by the trained system and sound source localization is achieved. The advantage of the learning-based approach is to be able to benefit from interaural level differences (ILD), which are not used in many applications since it is not possible to get a simple algorithmic solution showing the relationship between ILD and source location. Huang *et al.* developed a mobile robot auditory system consisting of an array of four microphones for sound localization and separation [7]. Their system depends on the onset detection of sound signals to decrease the negative effects of reflected sound. Many researchers have tried to imitate the sound localizing systems of animals that have strong sound localizing capabilities such as the barn owl [8].

PIR motion detectors are cheap and easy to use. In addition, processing infrared data is much easier than processing acoustic or video data. In video-based systems, it is not very practical to use more than two cameras (because of the computational cost), and two cameras cannot cover a large region. Since PIR are inexpensive and convenient to use, many of them can be used for covering a large area without increasing the computational cost considerably. Moreover, PIR detectors can work in the dark whereas vision based systems typically cannot. They can detect small temperature changes and hence can detect human body movement (since the temperature changes very little as a human passes through the field of view). A researcher from Naval Research Laboratory used passive infrared motion detectors to determine whether the obstacle is actually a human being [9].

The ultimate goal of the current robotics research is to develop robots that are capable of doing some of the activities that a human does, such as not only sensing but also perceiving their environment through audio and vision sensors. A robot that is supposed to behave autonomously is likely to need audio and visual sensors such as cameras and microphones. The microphones and cameras should imitate ears and eyes respectively and the robot should be able to turn its cameras to the direction of a perceived sound, as a human turns his head to face a sound source.

In this research, two passive human–robot interaction methods are implemented on a human service robot called ISAC. A sound source localizer (and human tracking) system using two electret microphones has been developed and placed on the upper part of ISAC (to imitate ears) for continuously localizing a person talking. This audio system is combined with the vision system of ISAC in order to make its cameras focus on the person and track him/her in real-time. Voice is one of the most natural human–robot interaction methods. This system can be integrated with a speech recognition system

to localize the human, and understand, to some extent, what the human says. Real-time tracking of the human makes this technique very useful for the service robot, since it can be used as a supporting tool for other passive interaction systems such as face tracking. Kismet, an autonomous robot developed at the Massachusetts Institute of Technology (MIT), Cambridge, [10], has similar active vision and audio systems consisting of two cameras (in the eyeballs) and two microphones (in the ears). Hadalay 2 is another humanoid robot developed at Waseda University, Tokyo, Japan, with audio-vision system [11].

In addition, a motion detector (and human tracking) system using an adjustable infrared sensor array consisting of five PIR sensors has been developed and placed on the body of ISAC. PIR sensors are sensitive to human body temperature changes and can be used to detect the movements of humans. This system is integrated with the vision system of ISAC in order to perform real-time human tracking. With the system developed and implemented on ISAC, it is possible to track two people moving in front of ISAC. Real-time human tracking with PIR detectors is another very useful and inexpensive passive human–robot interaction technique that can be combined with other methods such as face detection.

The present human tracking system of ISAC depends on a face tracker that can give unpredictable results for different conditions, e.g., lightening. In addition, humans interacting with ISAC by using face-tracking system should look at ISAC most of the time. Otherwise, ISAC loses the track of the face. Such systems are not fully passive and humans sometimes do not feel comfortable. In this study, two human localization and tracking systems depending on sound source localization and PIR motion detection have been integrated with ISACs face-tracker system to direct ISACs attention and prevent it from being quickly distracted. The integrated system increased ISACs current human-tracking capability by almost 40%. This increased the reliability of ISAC in many critical applications involving human interaction. Humans interacting with ISAC by using this system also feel more comfortable since they do not have to look at ISAC all the time. As the interaction gets more passive, the social acceptance of robot gets higher.

There are large number of applications of the integrated system, consisting of face-tracker, motion detectors, and sound localizer; not only in robotics but also in other areas such as automatic camera control, human–computer interaction, surveillance, monitoring, and entertainment.

This paper is organized as follows: Human–robot interaction categories and the differences between them are explained in Section II. Section III describes the hardware and software structures of ISAC. The human tracking systems based on sound source localization and PIR motion detection are explained in Sections IV and V, respectively. Section VI presents the integration of these systems and describes the experimental results. Finally, some conclusions are given and future work is motivated in Section VII.

II. HUMAN–ROBOT INTERACTION CONSIDERATIONS

Before diving into the details of the implemented systems, investigating Table I, which summarizes the interaction classifi-

TABLE I
CONSIDERATIONS FOR HUMAN–ROBOT INTERACTION

LOCATION	INTERFACE	FLOW OF NEW DATA	HUMAN EXPERIENCE OF ROBOT
Immediate: The robot is in the direct physical presence of the human.	Natural: e.g. voice and gestures.	None: e.g. human operates robot.	Human's Point of View: The human sees the world through his own eyes.
Intermediate: The robot is not in the physical presence of the human, but is not far away.	Machine: e.g. GUI, keyboard and joystick.	To Robot: e.g. autonomous robot exploring.	Robot's Point of View: The human sees the world through eyes of the robot.
Remote: The robot is far away from the human (with prior knowledge).		To Human: e.g. human explores using the robot.	3rd Party Point of View: The human sees the world through the eyes of another observer such as a camera attached on the ceiling.
Remote: The robot is far away from the human (without prior knowledge).		To Both	

cations of this research, is helpful for illustrating the differences between these modes. The main considerations are location, interface, flow of new data, and the human's experience of the robot.

A. Definitions

Immediate Location: The robot is in the direct physical presence of the human, and thus they share the same environment. The relative spatial relationships between the two are important. The human has direct visual contact with the robot, and a high bandwidth communication channel is likely to be available.

Intermediate Location: The robot is not in the physical presence of the human at all times, but is not far away. The spatial relationships are significant, but at a coarser level. The human may have some direct visual contact with the robot. A high or intermediate bandwidth communication channel is likely to be present.

Remote Location: The robot is far away from the human, and in an environment that is typically unknown to the human. The relative positions of the human and the robot are almost completely unimportant. The human does not have any direct visual contact with the robot. A low bandwidth communication channel is likely to be available.

Natural Interface: The human and the robot interact directly through natural means such as voice or gesture.

Machine Interface: The human and the robot interact through artificial means such as a computer keyboard or joystick.

In the light of Table I and the definitions given, the following concepts are developed and used throughout this paper. These do not cover all possible cases, but cover a few cases of interest to us.

Passive Interaction: The user does not need to behave in a specific manner but naturally. The robot has to be in the *immediate* location. It includes *natural interactions* such as via voice or gestures. New data may flow either to the robot only, or to both the robot and the human.

Semi-Active Interaction: This corresponds to the *machine interaction* and *human only learning* case. The human can learn

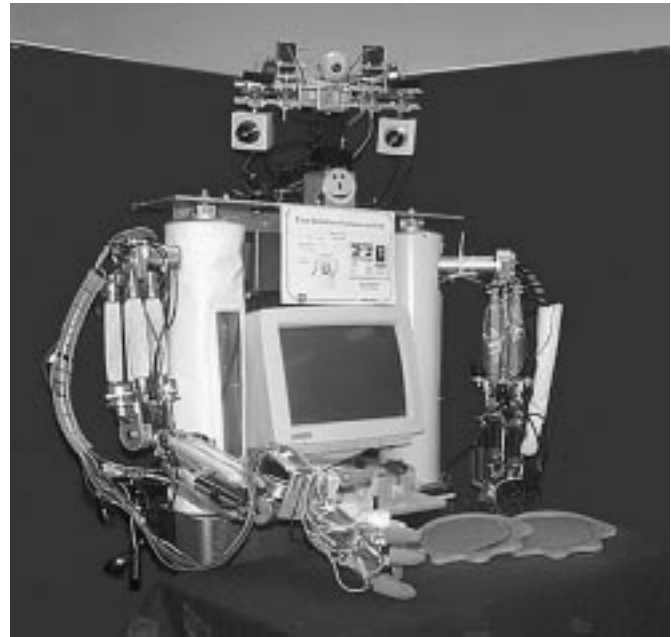


Fig. 1. Service robot ISAC.

about the environment the robot is in. However, the human does not teach the robot (in the sense of new data flow). The robot may be in the immediate, intermediate, or remote locations.

Full-Active Interaction: This corresponds to the *machine interaction* and both human and robot learning case. The human not only can learn from the robot but also can teach the robot (in the sense of new data flow). The robot may be in the immediate, intermediate, or remote locations.

In this research, two passive human–robot interaction techniques (human tracking based on sound source localization and PIR motion detection) have been implemented and integrated with the face-tracker system of ISAC to improve its existing human localization and tracking system.

III. DEVELOPMENT PLATFORM

ISAC, shown in Fig. 1, is a dual-arm humanoid robot used as a development platform in the Intelligent Robotics Laboratory, where it was designed and built.

A. Hardware Structure

ISAC has two 6-DOF arms, an active, stereo, color, vision system, 6-axis force-torque sensors at the wrists, simple haptic sensors on the fingers of its anthropomorphic hands, haptic sensors and proximity sensors on the palms of the hands, and ISAC has an array of infrared motion detectors on its torso. The robot also has two electret microphones for sound input.

B. Active Vision System

Vision is often necessary for robots that work with and around humans. The active vision system locates a face from stereo images and fixates upon it. Fixation is the process of centering each camera on a target point (a face in this example). This behavior is similar to that of a human turning his/her head as someone walks in front of him or her.

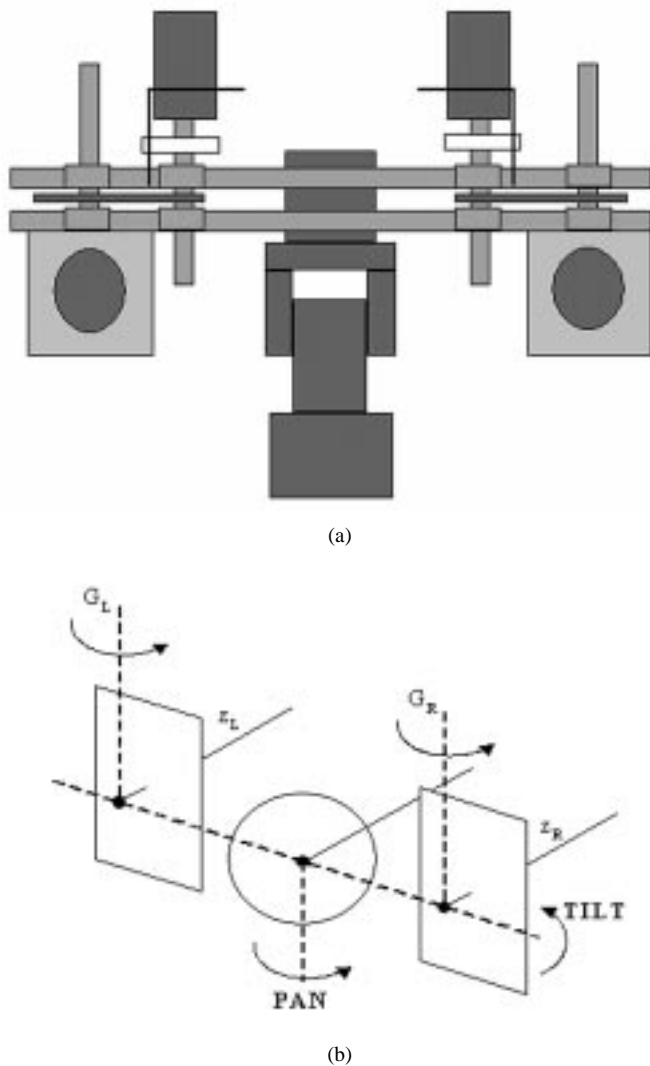


Fig. 2. (a) Four DOF camera head with two CCD color cameras and (b) axes of movements of pan, tilt, left verge, and right verge.

The active vision system of ISAC is composed of a four-DOF camera head (pan, tilt, left verge, and right verge) and two-color CCD cameras as shown in Fig. 2(a). Fig. 2(b) shows the movement axes.

The goal of the color module is to help locate a face in a set of stereo color images and guide the camera head to center it in both cameras. To accomplish this, skin tone color models were created. The color segmentation routine separates pixels into candidate skin-tone pixels and nonskin-tone pixels. Pixels are segmented by determining if they fall within a pre-defined RGB color model space. From this process a binary image mask is created.

Given the location of the skin-tone centroid in the mask image, the camera tracker then moves the camera head to guide the centroid toward the center of the image. The cameras move according to a direction vector generated by the distance from the skin-tone centroid to the center of the image view. The amount of the cameras' movement is proportional to the magnitude of the direction vector. The y -component of the direction vector controls the tilt motor and the x -component controls the verge motors. Once the target has reached the

center, the tracker is acknowledged that the target has been fixated and then stops moving camera head. Once a skin color blob is fixated upon we check for faces by using a template matching routine. A confidence measure is used to indicate the success of the fixation routine.

C. Software Structure

The intelligent machine architecture (IMA) is an agent-based software system that was initially developed for the humanoid robot, ISAC. IMA permits the concurrent execution of software agents on separate machines while facilitating extensive inter-agent communication. Within the context of IMA, an agent is one element of a domain-level system description that tightly encapsulates all aspects of that element, much like the concept of object in object-oriented systems. It has sufficient generality to permit the simultaneous deployment of virtually any robot control architecture, from Sense-Model-Plan-Act to behavior-based. IMA provides a two-level software framework for the development of intelligent machines. The robot-environment level describes the system in terms of a group of atomic software agents connected by a set of agent relationships (we use the adjective "atomic" to mean "primary constituent," the building blocks from which all compound objects are formed.) The agent-object level describes each of the atomic agents and agent relationships as a network of software modules called component objects. At the robot-environment level, IMA defines several classes of atomic agents and describes their primary functions in terms of environmental models, the robot itself (called the self agent), behaviors, or tasks developed for the robot. The atomic agent serves as a superstructure for everything the software knows or does relating to an element of the robot, a task, or the environment. Each IMA agent acts locally based on its internal state and provides a set of services to other agents through various relationships. IMA agents can and do exist at different levels of abstraction, from low-level hardware interfaces called hardware/resource agents, through behavior agents and task controllers called sequencer agents to high-level, autonomous, interactive entities called compound agents. IMA runs under Windows NT 4.0. The Distributed Component Object Model (DCOM) handles communication between atomic agents transparently.

COM objects can communicate to each other through interfaces. For example, component-A can invoke the methods of component-B through one of the interfaces of component-B. IMA components are created by using active template library (ATL) and each component supports some COM interfaces. Mainly, there are six subclasses of IMA components: mechanism, representation, engine (active component), manager, link, and relationship. Among these, mechanism components contain algorithms to process information, representation components are used to create special containers to hold data, and managers are mostly representation specific components and they are used to manage representations such as update them.

IMA components do not need to reside on the same computer; instead they can reside on different computers that are connected via a local area network. If two different components at different computers are needed to be used together, the agent



Fig. 3. Audio-visual sensors of ISAC.

locator takes care of this. The agent locator is on the one of the connected computers and it has a list of all registered components with their locations. Hence, one component can invoke the other component as if they were on the same computer.

IMA is used at IRL for developing modular, distributed robot software. Through IMA individually written software components are put together to form complex software agents [12]. To make ISAC more “human-friendly” a multi-agent system is developed [13] in which the interaction with the human is handled through a human agent. A more detailed explanation of the whole system and specifically, the human agent is given in [14]. Our study, contributes to improving the human agents alternatives.

IV. HUMAN TRACKING BY SOUND SOURCE LOCALIZATION

Sound localization has relevance for various applications such as aids for the physically disabled, monitoring devices, virtual reality interfaces, human–computer communications, and automatic video conferencing units. In particular, automatic camera pointing systems for robots are of interest in this research.

Surprisingly, some animals, such as bats and dolphins, use active sound localization different from passive sound localization. In active sound localization, the animal emits a sound and analyzes the echo to localize the target, and in passive sound localization, the animal analyzes sound emitted by the target itself. Ultrasonic transducers, which have a resonant frequency of 20 kHz or higher, are the most commonly used sensors for active sound localization. The transmitter part of these sensors produces an acoustic signal while the receiver part detects the returning echo.

A. Objective and System Configuration

Sound source localization has been used to achieve a variety of diverse goals such as automatically pointing a video camera at a talker in a conference room, or directing machine gun fire at a spinner [15]. In robotics, especially some mobile robot applications, such techniques have been implemented for target localization [7]. In this research, ISAC is equipped with a microphone array consisting of two microphones in order to localize the person it serves.

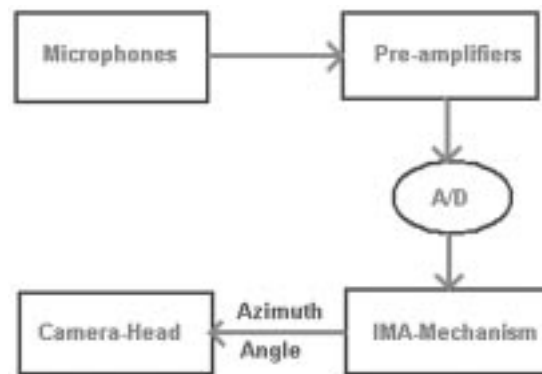


Fig. 4. Flow diagram of the procedure used.

A sensor configuration consisting of two small electret microphones is attached on the upper part of ISAC to localize a talker moving in front of it, as shown in Fig. 3. The outputs of the microphones are only several millivolts without any amplification. Hence, a pre-amplifier is used to increase the output voltage level of each microphone. The outputs obtained through the pre-amplifiers are sent to an A/D card to digitize the analog signals. The A/D card has four differential channels with a maximum aggregate sampling frequency 64 kHz. Since only two microphones are used, the maximum sampling frequency for each channel that can be achieved is 28–30 kHz. The sampled speech signals are processed by an IMA mechanism and the azimuth angle of the talker is calculated and sent to the camera head agent to rotate the camera head toward the talker by the calculated azimuth angle. The procedure is summarized in Fig. 4.

B. Time Delay Calculation

In the human auditory system, the head acts as a baffle between the two ears to affect intensity level of sound. Therefore, interaural level difference (ILD) may sometimes be a more important cue than time-delay for human sound localization. However, in this research, since the space between two microphones is open air, the effect of the head can be ignored and ITD (interaural time difference) can be considered as the most important cue. An illustration is shown in Fig. 5.

The sound source is assumed to be in front of the robot. Moreover, the two microphones are assumed to be on the same horizontal level. For the time delay estimation used in this research, the following notation is used.

$s(n)$	sound source waveform;
$s_1(n)$	left microphone output;
$s_2(n)$	right microphone output;
$n_L(n)$	noise for left microphone;
$n_R(n)$	noise for right microphone;
c^1	speed of sound;
d	separation of microphones;
d_L	distance between the sound source and left microphone;
d_R	distance between the sound source and right microphone;

¹ $c = 331.4\sqrt{(T/273)}$ m/s, where T is absolute temperature in Kelvin. At room temperature, $c = 343.2$ m/s.

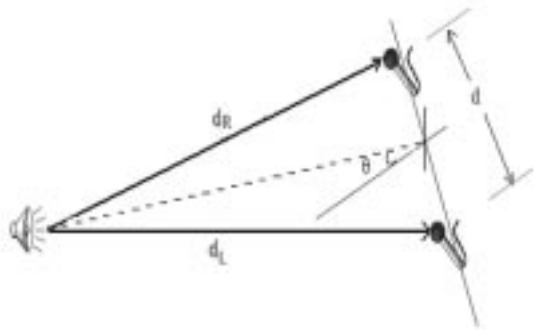


Fig. 5. Microphones and sound source configuration.

θ azimuth angle between the sound source and microphones.

When the sound source is far away from the microphones, the azimuth angle can be approximated as follows:

$$\sin(\theta) = (d_R - d_L)/d \Rightarrow \theta = \sin^{-1}(\tau \times c/d)$$

where τ is the time delay that needs to be estimated.

1) *Cross Correlation Method:* The cross-correlation (CC) method is used to estimate time by measuring the time at which cross-correlation of the left and right microphone waveforms reaches a maximum. It is perhaps the most basic and common method to estimate time delay, since it is a robust method for comparing signals originating from the same source. When the sound signals are obtained in an environment that is free of reverberations, and they are properly filtered, cross-correlation (in fact the generalized cross correlation) method reduces to the maximum likelihood time delay estimator and is asymptotically efficient in the limit of long observation times [16]. In reverberant environments, the delay estimations by cross-correlation are not very reliable since the sound signals are corrupted with echoes. Stephan and Champagne proposed a cepstral prefiltering technique to minimize the effects of reverberations [16]. Some researchers used redundant microphones to reduce the ambiguity occurring around the fundamental frequency of periodic waveforms [17]–[19]. Several other researchers also studied time delay estimation by cross-correlation under reverberant environments [20]–[23].

The noisy left and right microphone signals without multi-path distortion are modeled as

$$\begin{aligned} s_1(n) &= \alpha_L s(n - \tau_L) + n_L(n) \\ s_2(n) &= \alpha_R s(n - \tau_R) + n_R(n) \end{aligned}$$

where τ_L and τ_R are the time delays to reach to the left and right microphones, respectively, and α_L and α_R are scaling factors (which are inversely proportional to the distance between the microphones and the sound source) due to the intensity loss between the two microphones. According to the cross-correlation method, the value making the cross-correlation function of maximum gives the time delay as follows:

$$\begin{aligned} \tau &= \tau_L - \tau_R = \max_{\Delta} R_{12}(\Delta) \\ &= \max_{\Delta} \sum_{n=-\infty}^{\infty} s_1(n)s_2(n - \Delta). \end{aligned}$$

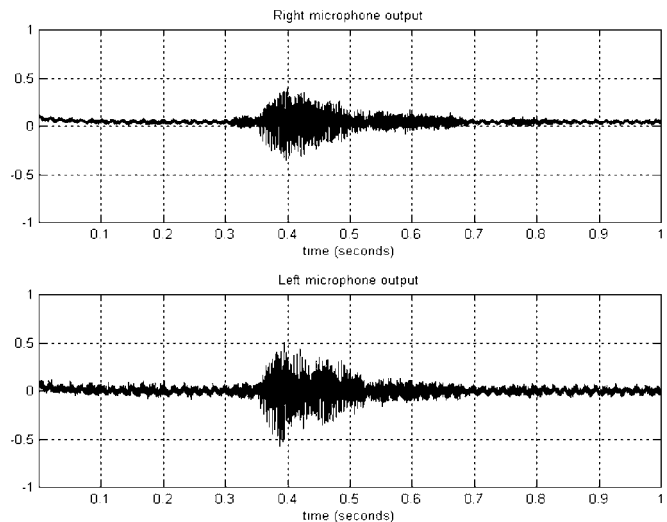


Fig. 6. Speech waveforms received by the left and right microphones when the sound source is on the left side.

In practice, the time delay is calculated as

$$\tau = \max_{\Delta} \sum_{n=-N}^N s_1(n)s_2(n - \Delta)$$

where N is the size of the window the cross-correlation is applied.

2) *Results:* In order to make an experimental analysis, the word “test” was spoken twice by a male talker in front of the microphones, once from the left side with the azimuth angle 35° and once from the right side with an angle of 28° . The length of the speech signals examined is 1.5 s. The speech waveform is partitioned into windows of 512 samples. For each window, the cross-correlation function of the left and right microphones is calculated and the maximum cross correlation value and the corresponding time delay are stored into a two dimensional array. After all windowed signals are processed, the time delay of the window with the maximum cross-correlation value gives the time delay of arrival between the left and right microphones.

Fig. 6 illustrates 1 s-long waveforms received by the left and right microphones, and a windowed portion of them is shown in Fig. 7.

The cross-correlation function of the left and right microphone outputs for the same time interval as in Fig. 7 is displayed in Fig. 8. Note that the index corresponding to the maximum value of the function is the time delay for that window. Also, the cross-correlation function corresponding to the whole signal is shown in Fig. 9. The maximum value for the whole signal has a maximum of 100 while it is only around 25 for the windowed portion. That is because the window chosen is not the window in which the cross-correlation has the maximum energy. The time delay corresponding to the maximum cross-correlation value over all the windows is -26 , which corresponds to a delay of 0.928 msec and an azimuth angle of 32.06° . The percentage error is 8.4%.

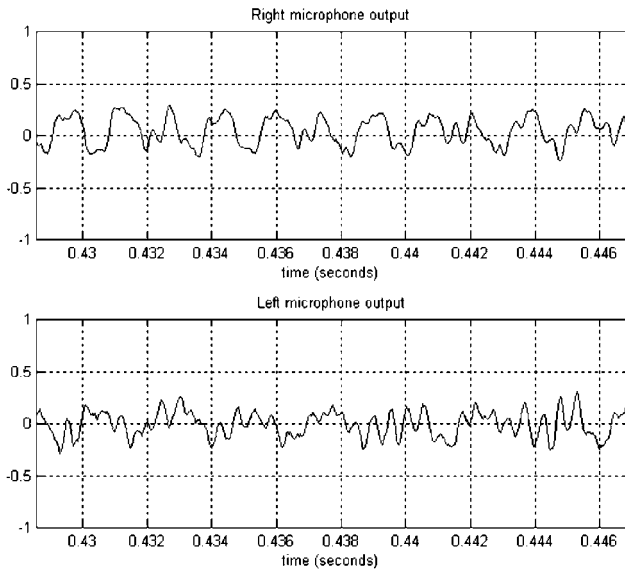


Fig. 7. Speech waveforms received by the left and right microphones (only 512 samples corresponding to 0.4286–0.4469 s).

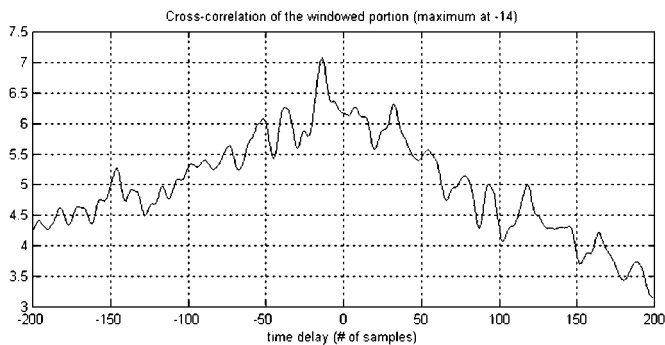


Fig. 8. Cross-correlation for the absolute value of the windowed portion.

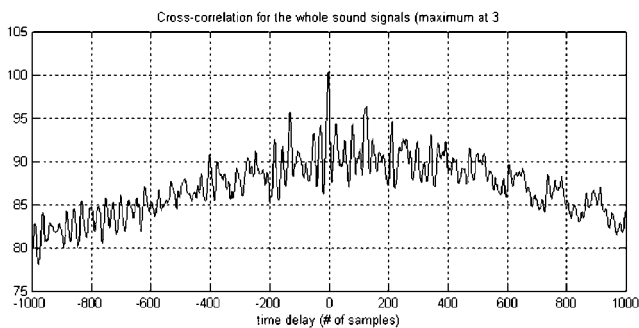


Fig. 9. Cross-correlation for the absolute value of the whole signal.

C. System Performance

In order to perform real time tracking with sound localization, an IMA mechanism (called the sound localizer) has been written. This mechanism reads data from the A/D card every 2 s and calculates the azimuth angle between the microphone array and the sound source. If the level of the cross-correlation is above a threshold level, it sends the calculated azimuth angle to the camera head component to make the cameras turn toward the detected person.

TABLE II
EXPERIMENTAL RESULTS FOR THE SOUND SOURCE LOCALIZATION

SOURCE LOCATION	AZIMUTH ANGLE (deg)	ESTIMATED AZIMUTH	ABSOLUTE ERROR
Left	35.00	32.06	2.94
Right	28.00	24.10	3.90
Left	16.00	19.43	-3.43
Right	12.00	9.60	2.40
Left	70.00	62.10	7.90

The system performance results are illustrated in Table II. The results are sensitive to the background noise level. Also, reverberations are another source of error. Experimental results have shown that the maximum error in azimuth estimate is less than 22%.

V. HUMAN TRACKING BY PIR MOTION DETECTION

A low-cost adjustable infrared sensor configuration consisting of five identical passive infrared (PIR) motion detectors has been added to ISAC. The intersections and unions of the motion detectors' outputs are used to track a moving object in three dimensions. Each motion detector is placed on a 5-in-long square piece of wood. The infrared motion detectors produce a 5 V output when they detect a moving object. Otherwise, an output of 0 V is produced. ISAC has a camera head containing two cameras. Four bits of information are extracted from the measurements. These are used as the tilt and pan angles of the camera head of ISAC and verge angles of the left and right cameras of the camera head.

A. Objective and System Configuration

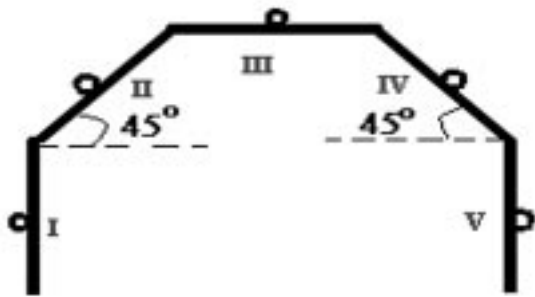
In this research, a sensor configuration consisting of five low-cost PIR motion detectors is placed on the body of ISAC to detect and track a person (or people) walking around. ISAC has two cameras placed on its pan-tilt unit. The overall vision system (cameras and pan-tilt unit) is placed on the head of ISAC. If there is a person walking in front of ISAC, both of its cameras look at the person.

The transducers in our infrared system are 7860-KT PIR detectors developed for human body detection. It can detect the movement of body heat in the horizontal direction, up to 190 cm away with nearly 180° field of view. The sensor configuration is shown in Fig. 10. Note that the angles between the wood pieces can be adjusted. Fig. 10(c) illustrates the sensitivity region of one of the detectors. Its detection region can be thought as a half ellipse with a horizontal length of 190 cm and a vertical length of 75 cm. Each detector produces 5 V when a human body moves in its range, otherwise the output is 0 V.

Each of the PIR motion detectors is connected to a channel of a Digital I/O card, and the digital output of each detector is then sent to an IMA mechanism (motion detector). This mechanism processes the digital outputs and calculates the coordinates of the detected person. It then finds the suitable pan and tilt angles of the camera head, and the verge angles of the right and left cameras. This procedure is illustrated in Fig. 11.

B. Tracking Methodology

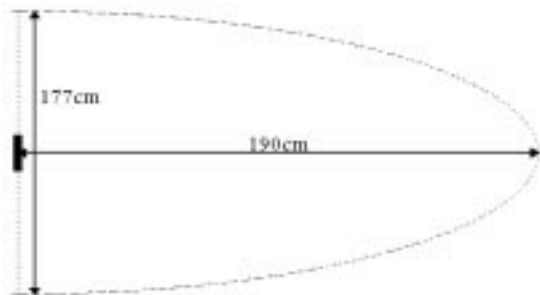
In this section, the methodology used to implement the real time multi-target tracking and camera pointing system is de-



(a)



(b)



(c)

Fig. 10. PIR sensor configuration (a) top view, (b) side view, and (c) sensitivity region of one of the PIR detectors.

scribed. Each detector has a sensitivity region in which it can detect as explained in the previous section. The tracking algorithm makes use of the intersections and unions of these sensitivity regions. Fig. 12 displays the intersections and unions of the sensitivity regions. The main idea is to decide which sensors can detect a human body in which regions.

Table III shows the regions and the transducers that are active in these regions. For example, in Region 7, the second, third, and fourth transducers are active and in Region 11, second, third, fourth and fifth transducers are active.

Note that the sensor configuration and the camera head are not on the same level. According to Table III, the pan angles of the cameras and the pan and tilt angles of the camera head are calculated. The camera head is pointed to the center of the region in which the person is detected. For example, when the person moves from Region 4 to Region 3, the pan angle of the camera head does not change whereas the tilt angle decreases. In some regions the pan angles of the right and left cameras are the

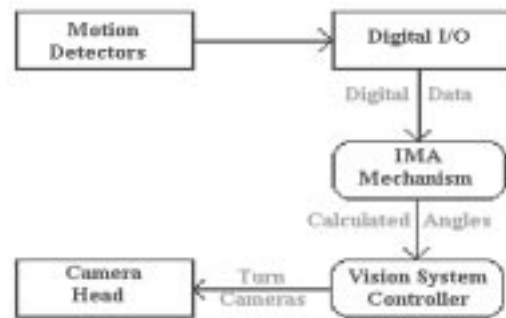


Fig. 11. Communication among the components.

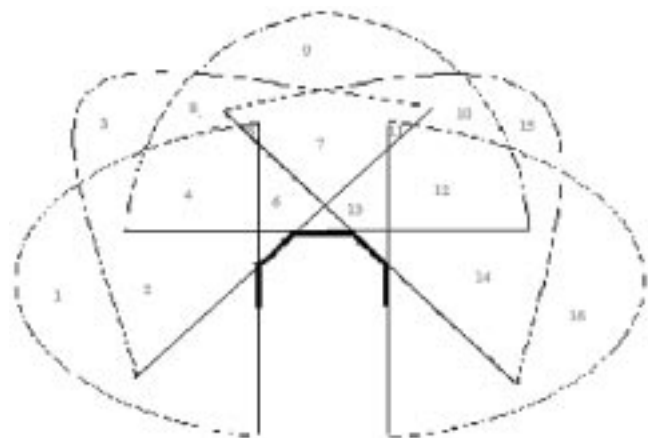


Fig. 12. Unions of the sensitivity regions.

TABLE III
ACTIVE DETECTORS IN SPECIFIED REGIONS

REGION	DETECTORS
1	I
2	I-II
3	II
4	I-II-III
5	I-II-III-IV
6	II-III
7	II-III-IV
8	II-III
9	III
10	III-IV
11	II-III-IV-V
12	III-IV-V
13	III-IV
14	IV-V
15	IV
16	V

same (e.g., Region 4) while in some regions they are different (e.g., Region 1). For example, for Region 1, the camera head is pointed to the center (by adjusting the pan and tilt angles of the camera head), the right camera is pointed to the upper section of Region 1, and the left camera is pointed to the lower section.

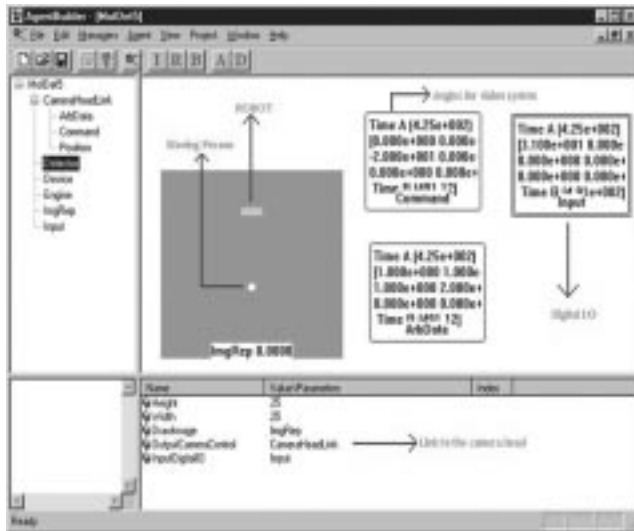


Fig. 13. Real-time human tracker agent.

C. System Performance

In order to implement the real time human tracking system with the adjustable PIR motion detector configuration, an IMA mechanism (called motion detector) has been developed. Digitized data from the Digital I/O card are taken by this mechanism every second and processed to localize a human (or humans) walking around. Moreover, a visual interface showing the path the person follows has been implemented as shown in Fig. 13. As the person moves, the white rectangular moves accordingly on the *ImgRep* as well. *Height* and *width* represent the height and width of the image representation respectively. *Input* represent the data coming from the Digital I/O card. For example, in this figure, Digital I/O reading is 31 which means that second, third, and fourth sensors are active. *Command* sends the suitable angles to the vision system. For example, in the figure, the pans for the camera head and cameras are zero while the tilt for the camera head is -20 (turning down). Each component is connected to another component. The order is as follows.

- 1) Input (Digital I/O output) \rightarrow InputDigitalIO (of Detector).
- 2) InputDigitalIO is processed in Detector and suitable angles are calculated.
- 3) Those angles are stored in OutputCameraControl.
- 4) OutputCameraControl \rightarrow Command (of CameraHead-Link).
- 5) Command sends the angles to the camera head.
- 6) DrawImage \rightarrow ImgRep.

The signals from the sensor configuration were sampled every second. The results of tracking one person were very good. The person was always in the view range of at least one of the cameras. The cameras tracked the person with 100% accuracy in all regions except for regions 5 and 11 of which the boundaries are not sharp. When there were two persons in the impossibility regions, the results were satisfying. One of the cameras looked at one person while the other looked at the other person. When the people were outside the impossibility regions, the results were not so good since there are some regions in which it is not possible to be able to track two persons. For example, if one of the

people is in Region 1 while the other is in Region 2, the two cameras are going to look at Region 2 since the first and second detectors are activated simultaneously.

VI. INTEGRATION WITH FACE TRACKER

Audio-vision integration is a natural skill for most animals and human beings. Information obtained by sound source localization is used to redirect their vision (and attention) systems to the localized areas (and objects). Vision allows the association of sounds with discrete objects in the world [24].

In the face-tracker system, the human face should always be in the field of view. Suppose the human turns his/her back to the robot and continues walking around. The face-tracker system will fail since it will not detect any faces. At this point, the motion detection system can be used to continue tracking until he/she turns his/her back. Also, the door of the laboratory is not in the field of view of the cameras. Hence face-tracker system cannot be initiated as soon as the human walks into the lab by itself. The motion detector configuration consists of five PIR motion detectors that are placed in such a way that they can cover a large area including the door. So, as the human walks in, the motion detectors can detect the body and the cameras can turn toward the human. In other words, the motion detection can be used to initiate the tracking system.

In this research, a tracker system that makes use of three modules has been developed. In this system, the face tracker is used as the main module and the other two modules (sound and infrared) are used to recalibrate face-tracker module when the following applies.

- 1) The face-tracker module fails. For example, ISAC can lose tracking the human and look toward the wall. In this case, the human can tell ISAC to look at him/her by saying "look at me ISAC."
- 2) The sound localization and motion detection modules produce an attention signal that can overcome that of the face tracking module's. For example, if there are two people, first one of them can attract ISACs attention for a while and then the other person can talk long enough to get ISACs attention.

A. Methodology

The infrared and sound localization technologies independently monitor for location, create detection events, and send decaying attention signals to the tracker. At a new detection, the corresponding attention signal is reset to its initial value. A summing junction is used to excite, by summing the input attention signals, and inhibit, by subtracting, the current tracking attention signal. If the sum is above the threshold value, the camera head moves to the new candidate region and looks there for a face to track. The tracker's attention signal is then reset to its initial value and begins its decay as well. If the output of the summer is below the threshold, the tracking remains in its current state (tracking or rest) and cannot be redirected until the current attention decays enough to allow the external signals to influence the resulting attention point. The output of the summer, before

thresholding, is the current attention signal $A_{\text{Track}}(n)$ that can be written as the sum of three terms

$$A_{\text{Track}}(n) = A_{\text{Sound}}(n) + A_{\text{Infrared}}(n) - A_{\text{Face}}(n)$$

Each term is given as follows:

$$\begin{aligned} A_{\text{Sound}}(n) &= A_{\text{sc}}^*(k_s)^{n-n_{\text{sa}}}u(n-n_{\text{sa}}) \\ A_{\text{Infrared}}(n) &= A_{\text{ic}}^*(k_i)^{n-n_{\text{ia}}}u(n-n_{\text{ia}}) \\ A_{\text{Face}}(n) &= A_{\text{fc}}^*(k_f)^{n-n_{\text{fa}}}u(n-n_{\text{fa}}) \end{aligned}$$

where $A_{\text{sc}}, A_{\text{ic}}, A_{\text{fc}}$ are constants for the sound, infrared, and face modules, respectively. $k_s, k_i,$ and k_f are exponential decay terms and $n_{\text{sa}}, n_{\text{ia}},$ and n_{fa} are the discrete time values at which attention signals are initiated by sound, infrared, and face modules respectively. The sum threshold, and initial and decay values combine to affect the system's sensitivity to particular stimuli or preference for certain events. This combination of factors determines how focused the system stays on a new task.

When $A_{\text{Tracker}}(n)$ becomes positive, it is considered that ISACs attention is distracted and the cameras look toward the point caused the attention and the face-tracker system is reset. Since $A_{\text{Face}}(n)$ is larger than $A_{\text{Sound}}(n)$ and $A_{\text{Infrared}}(n)$, the current attention is usually determined by the face-tracker system and it is mostly negative. However, after a while the signal coming from the face tracker will decrease and if a fresh signal comes from one or two of the other modules, then ISACs attention is distracted to another point.

B. Experimental Results

To validate this technology, we conducted controlled experiments to show that the tracking can be reliably initialized when a person enters the system's environment. The second aspect of the experimentation is during the tracking phase. The experiment will aim to show that the system's attentional focus can be switched between two people, with the switching cued by sound.

Three sets of measurements performed on ISAC to measure the improvement in tracking accuracy provided by the sound source localizer. Three 6-m long straight lines, which are 90 cm, 170 cm, and 240 cm away from ISAC respectively, were drawn in front of ISAC. First, only the face tracker is activated and a person was allowed to walk on each of these straight lines. The person walked slow enough to give enough computation time for the face tracker. If the person walks fast, the face tracker fails. In addition, the person first came very close to the cameras so that the face tracker detected his face and started tracking. Then, he walked away onto the lines drawn. The results are summarized as:

- 1) The face-tracker system failed in two trials out of eight when the person was on the first line. In addition, the face tracker failed whenever the person comes close to the edges.
- 2) It failed in five trials out of eight when the person was on the second line and it failed every time the person was close to the edges.

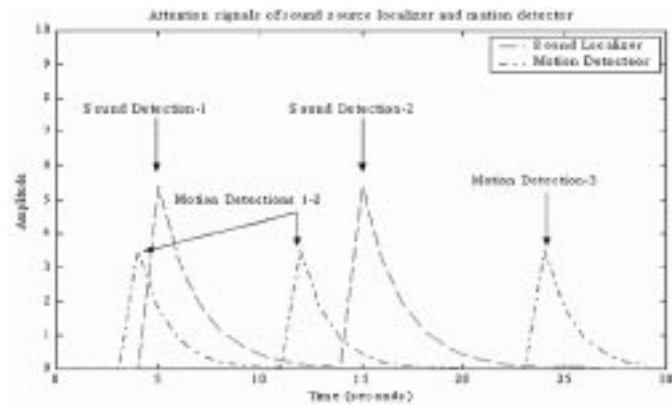


Fig. 14. Attention signals of the motion detectors and the sound localizer.

- 3) It failed in seven trials out of eight when the person was on the third line and again it failed every time when the person was close to the edges.

After that, the sound localizer was used to support the face tracker with an appropriate threshold value. The same person repeated the previous procedure, however, this time when the face tracker failed he spoke the phrase "Hello ISAC, I am here, can you look this way." The results are summarized as follows.

- 1) On the first line, the face tracker failed in three trials out of eight and the sound localizer corrected the system in one of them. In addition, it was able to locate the person correctly in one of the trials although the face tracker could not track the face.
- 2) On the second line, it failed in six trials and the sound localization corrected the system in two of them. Although the sound localizer was successful to take ISAC attention to the correct point one time, the face-tracker system could not continue tracking the face.
- 3) On the third line, the face tracker failed in seven trials out of eight and the sound localizer was able to correct the system in three of them. Again, the sound localizer system gained ISACs attention successfully in two times, however, the face tracker system could not continue tracking the face.

As a result, the integration of the sound localizer improved the tracking accuracy as follows.

- 1) On the first line, it was successful 66% and it improved the tracking system 33%.
- 2) On the second line, it was successful 50% and improved the system by 33%.
- 3) On the third line, it was successful 72% and improved the system by 42%.

In the experiments including the three systems, $A_{\text{sc}} = 7,$ $A_{\text{ic}} = 1.5,$ $A_{\text{fc}} = 10,$ $k_s = 0.6,$ $k_i = 0.3,$ and $k_f = 0.9$ values are chosen. Note that, since infrared sensors more frequently detect people that the sound localization, its decay factor is higher. Fig. 14 illustrates the attention signals produced by the sound source localizer and infrared motion detectors. Sound localizer detects some sound sources at times 5 and 15 s. Infrared detectors detect a motion at times 4, 12, and 24 s. Fig. 15 shows the overall attention system. The face-tracker system is distracted four times at 5, 12, 15, and 24 s. Otherwise, the face tracker is

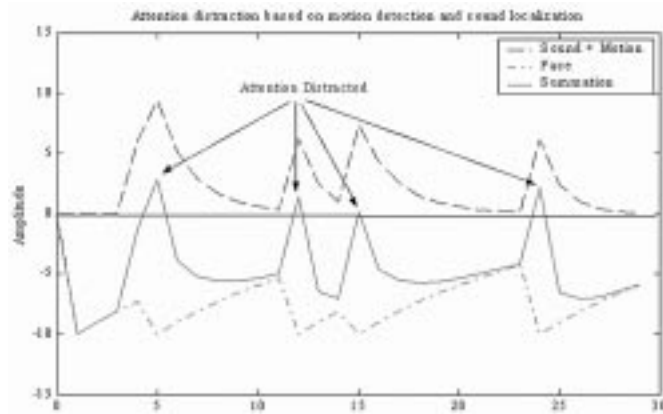


Fig. 15. ISACs attention distraction.

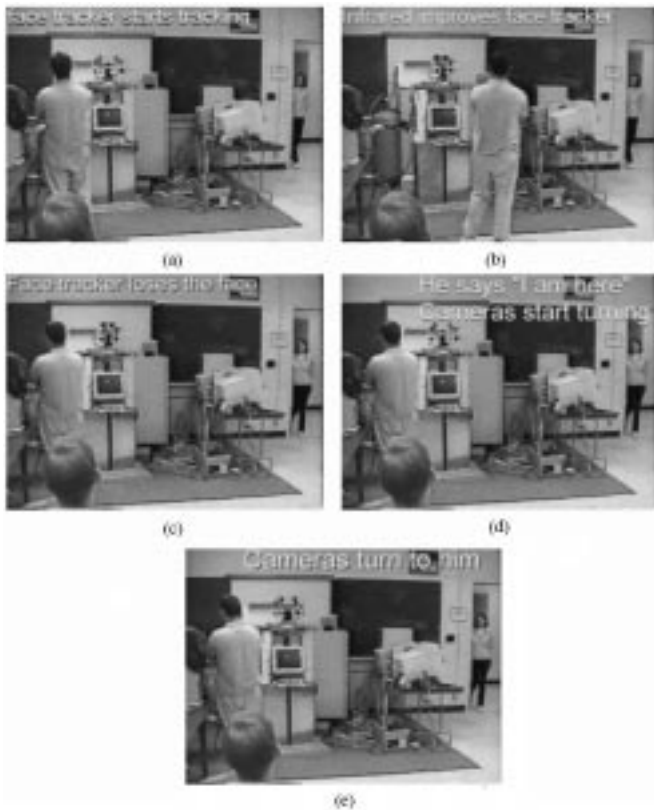


Fig. 16. Integrating face tracker, sound localizer, and motion detectors.

dominant over the sound source localizer and motion detector combination. Fig. 16 shows some images from the experiment performed.

The main objective of implementing passive human–robot interaction systems is to make human feel more comfortable with the robot. As the interaction requires less amount of human effort, people interacting with robots can act more naturally. When the face-tracker system used alone, human was required to look at ISAC almost all of the time so that the face-tracker system would not lose the face it is tracking. This could make people interacting with ISAC uncomfortable. The integrated system prevented this and people were able to act more naturally and they did not have to look at ISAC all the time. The system was able to track people (by detecting motion and localizing the sound source) even when they are not aware.

VII. CONCLUSION

In this research, ISAC is used as a test-bed to develop and implement some human–robot interaction techniques (referred to earlier as passive interactions). The interaction is passive in the sense that the user does not help the robot (i.e., the human does not operate the robot), he/she just behaves naturally. Two passive human–robot interaction implementations have been introduced and integrated with the face-tracker system of a humanoid robot, ISAC:

- automatic camera control/human tracking by sound source localization;
- automatic camera control/human tracking by human motion detection.

The results show that the module consisting of face tracker, sound source localizer, and infrared motion detector provides a robust human tracking system that imitates humans' attention distractions. This is the first time that the three systems: face tracking, sound localization, and infrared motion detection are combined with an attentional measure to develop a reliable tracking system.

REFERENCES

- [1] J. Borenstein, D. Webe, L. Feng, and Y. Kore, "Mobile robot navigation in narrow aisle with ultrasonic sensors," in *Proc. 6th Topical Meeting on Robotics and Remote Systems*, Monterey, CA, 1995.
- [2] M. Omologo and P. Svaizer, "Acoustic source location in noisy and reverberant environment using CPS analysis," in *Proc. ICASSP*, Atlanta, GA, 1996.
- [3] C. Wang and S. Branstein, "A hybrid real-time face tracking system," in *Proc. ICASSP*, Seattle, WA, 1998.
- [4] D. E. Sturim, M. S. Brandstein, and H. F. Silverman, "Tracking multiple talkers using microphone-array measurements," in *Proc. ICASSP*, Munich, Germany, 1997.
- [5] H. Wong and P. L. Chu, "Voice source localization for automatic camera pointing system in videoconferencing," in *Proc. Workshop on Applications of Signal Processing to Audio and Acoustics*, 1997.
- [6] K. Y. Guentchev and J. J. Weng, "Learning-based three dimensional sound localization using a compact non-coplanar array of microphones," in *Intelligent Environments Symp.*, Stanford, CA, 1998.
- [7] J. Huang, N. Ohnishi, and N. Sugie, "Building ears for robot: Sound localization and separation," in *Proc. Int. Symp. Artificial Life and Robotics*, Oita, Japan, Feb. 1996.
- [8] M. Rucci and J. Wray, "Binaural cross-correlation and auditory localization in the barn owl: A theoretical study," *Neural Networks*, vol. 12, pp. 31–42, 1999.
- [9] AAAI-97 Robot Competition [Online]. Available: <http://www.aic.nrl.navy.mil/~schultz/research/coyote/index.html>
- [10] C. Braezael, "A motivational system for regulating human–robot interaction," in *Proc. AAAI*, 1998, pp. 54–61.
- [11] S. Hashimoto, H. Kasahara, A. Takanishi, S. Sugano, K. Shirai, T. Kobayashi, H. Takanobu, T. Kurata, K. Fujikawa, T. Matsuno, T. Kawasaki, and K. Hoashi, "Humanoid robot—Development of an information assistant robot hadaly," in *Proc. IEEE Int. Workshop Robot Human Communication*, 1997, pp. 106–111.
- [12] R. T. Pack, "IMA: The Intelligent Machine Architecture," Ph.D. thesis, Vanderbilt Univ., Nashville, TN, 1998.
- [13] W. A. Alford, T. Rogers, D. M. Wilkes, and K. Kawamura, "Multi-agent system for a human-friendly robot," *Proc. 1999 IEEE Conf. Systems, Man, and Cybernetics (SMC'99)*, 1999.
- [14] K. Kawamura, R. A. Peters II, D. M. Wilkes, W. A. Alford, and T. E. Rogers, "ISAC: Foundations in human-humanoid interaction," *IEEE Intell. Syst. Mag.*, vol. 15, pp. 38–45, Apr. 2000.
- [15] D. V. Rabinkin, R. J. Renomeron, and F. C. French, "Estimation of wave-front arrival delay using the cross-power spectrum phase technique," in *132nd Meeting Acoustical Society of America*, Honolulu, HI, 1996.
- [16] A. Stephenne and B. Champagne, "A new cepstral prefiltering technique for estimating time delay under reverberant conditions," *Signal Process.*, vol. 59, pp. 253–265, 1997.

- [17] U. Bub, M. Hunke, and A. Waibel, "Knowing who to listen to in speech recognition: Visually guided beamforming," *Proc. IEEE Int. Conf. Auditory Speech and Signal Processing*, vol. 1, pp. 848–851, 1995.
- [18] H. F. Silverman and S. E. Kirtman, "A two-stage algorithm for determining talker location from linear microphone array data," *Comp. Speech Lang.*, vol. 6, no. 2, pp. 129–152, 1995.
- [19] K. Takahashi and H. Yamasaki, "Audio-visual sensor fusion system for intelligent sound sensing," in *Proc. Int. Conf. Multisensor Fusion and Integration for the Intelligent Systems*, 1994, pp. 493–500.
- [20] Y. T. Chan and P. C. Ching, "Non-stationary time delay estimation with a multipath propagation," *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, pp. 2736–2739, 1989.
- [21] P. C. Ching, K. C. Ho, and Y. U. Chan, "Constrained adaptation for time delay estimation with multipath propagation," in *IEEE Proc. Radar Signal Process.*, vol. 138, 1991, pp. 453–458.
- [22] M. Omologo and P. Svaizer, "Acoustic event localization using crosspower-spectrum phase based technique," in *Proc. ICASSP*, 1994.
- [23] J. O. Smith and B. Briedlander, "Adaptive multipath delay estimation," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-33, pp. 812–822, 1985.
- [24] S. G. Goodridge, "Multimedia Sensor Fusion for Intelligent Camera Control and Human-Computer Interaction," Ph.D. thesis, North Carolina State Univ., Raleigh, NC, 1997.



Ali Şafak Sekmen (M'00) received the B.S. and M.S. degrees in electrical and electronics engineering from Bilkent University, Ankara, Turkey, and the Ph.D. degree in electrical engineering from Vanderbilt University, Nashville, TN.

He is currently an Assistant Professor of Computer Science, Tennessee State University (TSU), Nashville. Previously, he was an Assistant Professor of Electrical and computer engineering at TSU. He has published over 40 research papers in robotics, intelligent systems, and signal processing. He was a

Member of the Intelligent Robotics Laboratory, Vanderbilt University between 1997–2000.

He has been involved in research projects including human-robot interaction, intelligent systems, mobile robots, humanoid robots, and component-based software systems development.



Mitch Wilkes (M'90) was born in Dallas, TX. He received the B.S.E.E. degree from Florida Atlantic University, Boca Raton. He received the M.S.E.E. and Ph.D. degrees in electrical engineering from Georgia Institute of Technology, Atlanta, in 1984 and 1987, respectively.

He is an Associate Professor of electrical engineering and computer engineering, School of Engineering, Vanderbilt University, Nashville, TN. He is also an Assistant Director of the Center for Intelligent Systems and the Assistant Director for

the Intelligent Robotics Laboratory. His research interests include intelligent robotics and control, signal processing, image processing, and intelligent manufacturing.

Dr. Wilkes currently serves as a member of the Technical Committee on Service Robots for the IEEE Robotics and Automation Society.



Kazuhiko Kawamura (M'69) received the B.E. degree from Waseda University, Tokyo, Japan, the M.S. degree from the University of California, Berkeley, and the Ph.D. degree in electrical engineering from the University of Michigan, Ann Arbor.

He is currently a Professor of electrical engineering and computer engineering, and Professor of management of technology, Vanderbilt University School of Engineering, Nashville, TN. He is also the Director of Vanderbilt's Center for Intelligent Systems and Intelligent Robotics Laboratory. He

has published over 120 research papers, a book, and book chapters in the fields of intelligent systems, intelligent robotics, intelligent manufacturing, and environmentally sound manufacturing. Previously, he was a Lecturer with the University of Michigan, Ann Arbor, and a Systems Planner with Battelle Columbus Laboratories, Columbus, OH. He directs research projects in intelligent systems, computational intelligence, humanoid robots, mobile robots, and human-robot interfaces.

Dr. Kawamura was a member of SMC AdCom from 1998–2001 and General Chair for the IEEE SMC 2000 Conference. He is Founding Chair of the Technical Committee on Service Robots for the IEEE Robotics and Automation Society. He also serves as a board member of the Academic Coalition for Intelligent Manufacturing Systems (A-CIMS).