Towards Reliable Multimodal Sensing in Aware Environments

Scott Stillman and Irfan Essa
GVU Center / College of Computing
Georgia Institute of Technology
Atlanta, Georgia 30332-0280, USA
sstil@cc.gatech.edu, irfan@cc.gatech.edu

ABSTRACT

A prototype system for implementing a reliable sensor network for large scale *smart environments* is presented. Most applications within any form of smart environments (rooms, offices, homes, *etc.*) are dependent on reliable who, where, when, and what information of its inhabitants (users). This information can be inferred from different sensors spread throughout the space. However, isolated sensing technologies provide limited information under the varying, dynamic, and long-term scenarios (24/7), that are inherent in applications for intelligent environments. In this paper, we present a prototype system that provides an infrastructure for leveraging the strengths of different sensors and processes used for the interpretation of their collective data. We describe the needs of such system, propose an architecture to deal with such multi-modal fusion, and discuss the initial set of sensors and processes used to address such needs.

1. INTRODUCTION

Building environments that support our daily activities has become an important research effort in recent years [24, 13, 5, 16, 20, 2]. Such environments include rooms, offices, automobiles, homes, *etc.* For applications beyond simple automation, these environments require both an explicit interface and an implicit inference engine to allow for effective interaction with inhabitants/users. An explicit interface in such environments requires development of off-the-desktop types of interaction (*e.g.,* speech and gestures), while an implicit inference engine needs to reliably determine *who* the user is, *where* the user is, *what* is being attempted, and *when*. Both these explicit and implicit interfaces need to reliably work over extended periods.

Several applications are under development that require very specific types of "awareness" on the part of the environment. For example, the Digital Family Portrait can non-invasively and privately monitor and report the level of daily activity for an elderly parent to a trusted group [18]. The Family Intercom needs to know where family members are and what they are doing [19]. It is particularly important that context about the status of the callee be communicated to the caller, so that the appropriate social protocol for continuing a conversation can be performed by the caller. An adult son may want to phone his mother living in another city to see how she is doing, but he does not necessarily want to wake her up from a nap in the process. In an ideal world, one would prefer not to have to resort to those measures. Having a fully "aware" environment of where its residents are with a high degree of reliability and a building that knows who is in what room are therefore significant capabilities required to provide many of the services being considered.

In this paper, we are also primarily dealing with a system that leverages off rich multimedia streams (specifically audio and video processing and interpretation). Our long-term interest is to merge these with other forms of sensing. We are also emphasizing passive sensors within the environment and not worn on the person. We expect those also to merge in the later part of the project.

Passive sensing technologies are far from meeting the requirements of robust tracking in largely dynamical, and changing environments, over extended time frames. In isolation, each passive sensing technology that we examined cannot satisfy the above requirements stated. Auditory localization, for example, under the best of circumstances gives the location of a speaker but not their identity. Under realistic and continuously changing conditions, dependable accurate readings of a speaker's location are much less certain.

The goal of this paper is to take some of the lessons learned in the computer vision, audio, and sensor fusion communities, and apply them to the design of a infrastructure that is robust, reliable, and near real-time. All of this done while providing quality data and reporting with regard to where people are, what they are doing and when they are doing it. This paper presents a system that provides a foundation for providing such services. The basic paradigm underlying our work on merging and fusion of different modalities follows our belief that exploiting a large amount of evidence from a variety of sources can be merged and reinforced by model knowledge through a probabilistic framework [26]

1.1 Related Work

There is much work in the area of multimodal fusion in the military domains. However, only recently have such efforts been considered for smart environments and for perceptual

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
Copyright 2001 ACM X-XXXXX-XX-X/XX/XX ...$5.00.

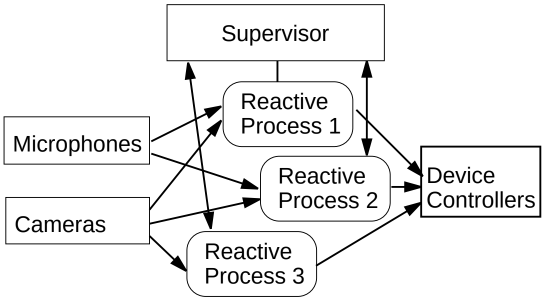


Figure 1: *A supervisory controller selects and controls the sequencing of perceptual processes. Multiple processes can be active at the same time.*

user interfaces. We describe some of the contributions that are closely related to our effort. Crowley and Bedrune [6] introduced the concept of an ensemble of visual processes controlled by a supervisor (graphically shown in Figure 1). The system is based on an architecture in which a supervisor selects and activates visual processes in a cyclic manner. Confidence factors accompany each observation and are used for reasoning with and controlling each visual process. In the figure, each reactive process receives perceptual input from a set of virtual sensors[1], shown here as microphones and cameras, and produces commands for driving a set of device controllers. The processes are organized as a network of states where each state corresponds to a set of reactive processes with associated control parameters. Multiple states can be active at the same time, and state transitions can be conditioned on unexpected events.

Crowley and Berard later used this framework to detect and track faces for a video communications system [7]. Blink detection (of the human eye), color detection, and correlation were the system's reactive processes that were used in controlling camera movement. Our system makes use of this concept of a supervisory controller for managing its own sensor resources.

Goodridge and Kay [14] provided another example of sensor fusion where acoustic and visual data are combined. A multi-sensor-based system has been developed for controlling a Pan/Tilt/Zoom (PTZ) camera for the purpose of tracking a speaking person. A skin-tone detection algorithm is used to identify skin pixels and to drive the determination of face location. In addition, two microphones are used for performing auditory localization. These two features from the audio and vision process are fused together to provide a measure of certainty as to the validity of an object. The data is fused together at the pixel level rather than the symbolic level to improve the detection of the speaker's face. In the fusing stage, the sound localization histogram is normalized and mapped to pixel coordinates. From that a conditional probability $P($ face at pixel i | sound $)$ may be generated and combined with $P($ face at pixel i | color $)$ to form a joint posterior probability. This metric is used for image segmentation to extract the face of the speaker from the image. This work in conjunction with [21] motivated the audio work described in the next section.

1.2 Our Approach

Sensors can be divided into the following classes [11]: complementary, competitive, and cooperative. Complementary sensors do not depend on each other directly but can be merged to form a more complete picture of the environment. Competitive sensors each provide equivalent information about the environment. A general problem with these kind of sensors involves interpreting conflicting readings. Finally, cooperative sensors work together to derive information that neither sensor could provide [4]. This type of fusion is dependent on details of the physical devices involved (e.g., microphones and cameras).

Our system, as described in the next section (Figure 2), is composed of a 3 layer hierarchal model. The first (lowest layer) is the Process Supervisor layer that uses a combination of complementary and cooperative sensors. For example, the cameras that monitor the doorway to a room act as complementary sensors–one camera monitoring the door for occupants walking in, and another more obliquely angled camera monitoring for people leaving the room. The second and third layer representing the room manager level and the house level use the outputs of the layer below its own as competitive sensors that provide who, when, and where information.

The goal of this work is to leverage the inherent strengths of different sensor types and algorithms applied to interpretation of collected data for the express purpose of obtaining reliable state information about residents within the home. In this paper, we prototype a system that provides an infrastructure for doing just that. Our initial experimentation has focused only on the use of cameras and microphones but other devices can easily be incorporated. All of this work is now in place in one of the rooms in the Georgia Tech's Broadband Institute Residential Laboratory, one of the centers of our research on Aware Homes. Further integration of this prototype with other sensing and processing approaches is underway. Our initial results show that using a combined sensor approach can improve the overall accuracy of state information of the house.

2. SYSTEM DESIGN

2.1 Audio and Video Sensors

Microphones have been set up in various configurations and serve different purposes throughout the space. In addition to performing traditional communication functions, microphones have been set up for auditory localization, speaker identification, and steered beam former arrays for speech recognition. Our system in its current state has focused on the localization capability with the other capabilities to be added later.

Of the various methods for performing auditory localization, time delay estimation based techniques have shown to be more effective in our real time environment based on [15, 21, 10]. An ideal free-field model was used and the model parameters were estimated using a phase transform method (PHAT). This method was chosen because it is simple, easy to implement, and provides a good response time, but will fail with excess reverberation or when the fundamental model assumptions are violated.

The term *video sensor* refers to the abstraction of applying a computer vision algorithm to the stream of video data captured by one of the many cameras in the house.

[1] A virtual sensor is defined as a time sampled function, $S_i(t)$, that is computed on a subset of the set of transducers $T_i(t)$ and intermediate representations $R_i(t)$.

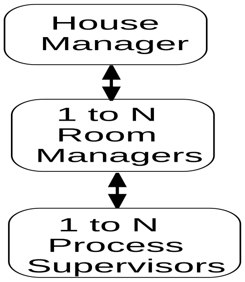


Figure 2: *A three layer hierarchy of Sensor Process Supervisors, Room Managers, and a House Manager provide the frame work for developing a reliable comprehensive system for delivering who, what, and where information for Aware Home applications that require it.*

Various vision processes applied to this data allow for the creation of different hypothesis that can be tested or verified with other sensor data. Tracking methods, background differencing techniques, and face recognition algorithms are the types of processes being applied. The visual processes implemented currently focus on the need for occupant identification and their location.

The assumption most often violated in auditory localization is that the average energy of the captured auditory signal is significantly greater than that of the interfering noise. Vision sensors enable complimentary analysis that helps minimize the impact of invalid assumptions when they occur.

2.2 Hybrid Supervisor/Manager Architecture

Now that we have a variety of sensor devices and processes to supply data, we need a means for aggregating data, making decisions, and taking appropriate actions. Various types of server models were looked at including synchronous, multi-threads, and asynchronous servers. The Sensor Process Supervisor (SPS) configuration is shown in Figure 1. Here all sensor processes (e.g., face rec., auditory localization, etc.) feed into a central process supervisor. The function of the SPS is to collect data from complementary and cooperative sensors, process the data, and send updates to the camera control processes that drive the PTZ cameras.

Some sensor processes can be combined to form a process that is capable of performing more than one method or algorithm. This is useful for cases where you don't want to dedicate a particular camera to perform face recognition solely. The SPS can choose which method or algorithm it deems most useful for helping analyze its current data. Such a multiplexing scheme appears to be a useful approach in our work so far.

After processing a sample period worth of data, the SPS sends an update to the Room Manager (RM) regarding what "significant" events it believes have occurred. The RM takes this data and updates its view (state) of the room and stores the information in an occupancy or evidence grid as stated by [17]. Examples of significant events are "Room Occupancy = 3" or that a particular occupant is speaking at a certain location with a certain level of confidence. There is only one RM per room and multiple SPSs per room. The RM uses a competitive sensor model to process the information it gets from the its SPSs.

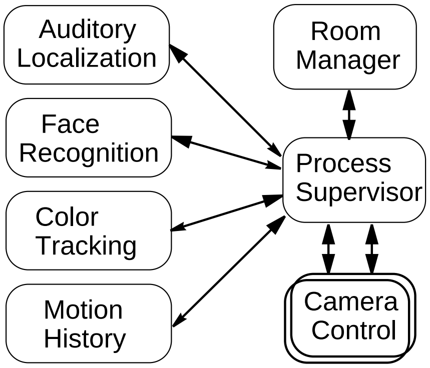


Figure 3: *The Sensor Process Supervisor collects and analyzes information from various sensor processes. Upon completion of this task, it updates its occupancy grid and sends any event data it deems to be significant to the Room Manager.*

Both vision and audio sensor placements are calibrated in room coordinates. The SPS uses these measurements to report identified targets within a common framework to the RM. Raw data collected by the sensors typically does not go beyond the scope of the SPS.

A Bayesian approach for combining sensor data at the RM as outlined by Moravec and Blackwell[17] was considered. This framework, however, relied heavily on the assumption that individual sensor readings were independent, which is clearly not always the case. Cells within the occupancy grid were instead populated with normalized probabilities based on the sensor type, confidence measures, plausibility measures, etc.

Given the implementation of multiple SPSs per RM and one RM per room, the next layer of integration would be to implement a House Manager(HM), as shown in Figure 2. This would allow for complete end-to-end monitoring of the house. A future research goal is to have the RM and the HM wrapped into the Context Toolkit Framework [9].

An aware environment requires historical data detailing previous movements of its occupants. The Room Manager is responsible for recording these events to a file for later use and also for updating the appropriate context toolkit widgets. Names of occupants are recorded if one of the identifier processes (face recognition or speaker identification) is successful. In addition, their locations are time stamped and recorded along with corresponding confidence levels.

3. CORRESPONDENCES AND COMBINATIONS

3.1 Sensor Configuration

Figure 4 shows a top-level view of one of the rooms with direction North as indicated. Two cameras, one on the south wall and one on the west wall have been dedicated to the task of face detection and recognition. The camera on the west wall remains stationary and focused on the entrance of the door located on the east wall. This gives the system its best chance at performing a correct face recognition as an occupant enters the room. The camera on the north wall

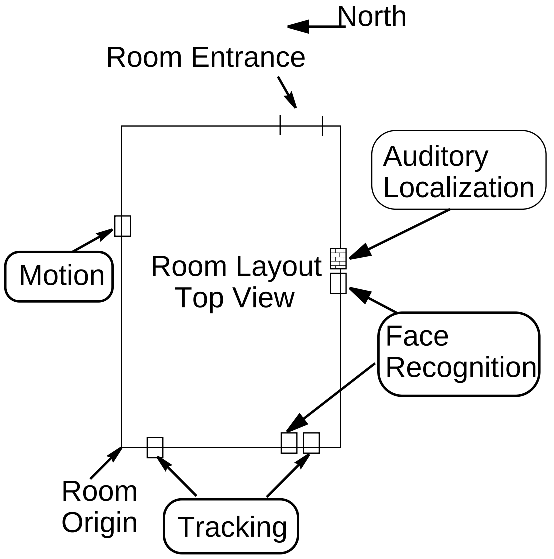


Figure 4: *This layout of the room has 5 cameras and 4 microphones. The microphones are configured in a 20 by 20 cm square and are used to perform auditory localization in 2 dimensions. The cameras are configured to perform the functions shown.*

is a Pan-Tilt-Zoom camera but stays mainly focused on the doorway entrance. This is the primary sensor for detecting a person leaving.

The PTZ cameras are Canon VC-C3 cameras and are controlled by a serial interface. Each camera is aligned such that its center position is pointing directly perpendicular to the opposite wall. Angles to targets can then be estimated by reading its current position as shown in Figure 5. Using the zoom feature of the VC-C3 will not affect this angle.

3.2 The Sensor Process Supervisor

The Sensor Process Supervisor (SPS) is a process that takes input from the various sensor processes and provides complementary and cooperative fusion of the data. An SPS can support both at the same time but for simplicity, we configured two SPSs: one for complementary sensors and one for cooperative sensors.

The case for cooperative sensing is where we have assigned the two cameras on the west wall to perform tracking tasks. Each camera process reports a list of potential objects(occupants) to its SPS along with corresponding feature information. The SPS then performs a matching function to make the correspondence [23]. Typical features include centroid coordinates, velocity, color statistics, and timestamps. Once an occupant has been matched, the position of the occupant is calculated for the x direction given the angles α and β and the base length (L) between cameras as shown in the figure 6. The calculation is:

$$x_{occupant} = L * \frac{tan(\alpha)}{tan(\alpha) + tan(\beta)}. \tag{1}$$

From that the occupant y direction can be calculated by:

$$y_{occupant} = \frac{x}{tan(\alpha)}. \tag{2}$$

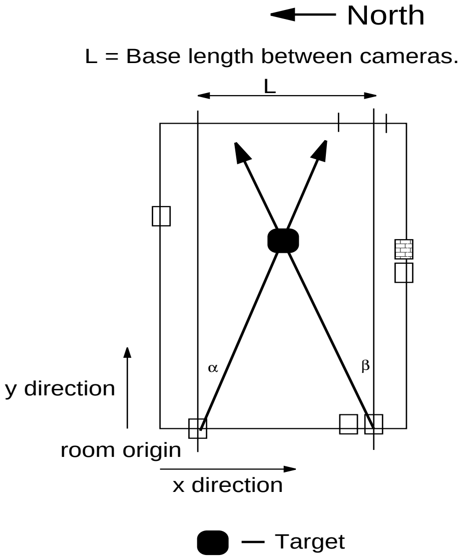


Figure 5: *Triangulation of the object angle data is used by the SPS to estimate the position of occupants.*

The case of the complementary sensing is when the face tracker from the visual sensor is combined with the audio-based speaker tracker to ensure the location of a speaker in the space.

3.3 The Room Manager and Occupancy Grids

We have decided upon the usefulness of loosely categorizing the occurrence of events as: *instantaneous, recent past, and archived*. Archived data is considered data that gets logged for things like data warehousing and On-line Analytical Processing (OLAP). *Recent past* events are used for developing hypotheses about residents as they move about the house. *Recent events* will be archived unless they are found later to be erroneous, in which case they are removed from the log. Archived and recent past events are deemed *significant* events.

Instantaneous events are used to formulate significant events but are not necessarily significant themselves. All the sensors report instantaneous events with no regard for their importance. We have not formalized methods for such determinations, but realize the importance of such. Crowley and Demazeau discuss the need for the process of 'intelligent forgetting' so as to prevent the internal model from growing without limits[8].

We are looking first at keeping track of residents as they move about the home. Our system specifically focuses on tracking a person from the time they enter a room until the time they leave the room.

The Room Manager (RM) is the process that is responsible for tracking significant events (namely the who, where, and when of the occupants). It takes input from different SPSs and formulates the state of the room. The RM uses an evidence grid to capture the spatial knowledge of events that occur in the room as shown in Figure 7. The work of Elfes, Moravec, and Blackwell have shown that this approach allows for the efficient accumulation of small amounts of information from individual sensor readings into an increasingly accurate and confident map of the environement [12, 17].

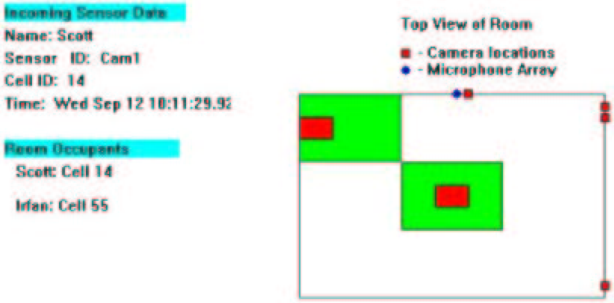
4. RESULTS

The initial configuration has two Sensor Process Supervisors (SPS) and one room supervisor. One SPS uses two cameras as shown in Figure 6 to track occupants as they enter the room. Initial analysis was done to determine the accuracy of using the angle calculations. Test objects were placed at 15 different somewhat random test locations throughout the room. The angles were estimated using the process outlined in this paper and compared against the actual (measured) angles. On average, our results were within 10 cm of the target. The CAMSHIFT algorithm is used to perform occupant tracking with both cameras [3]. Our initial implementation only uses centroid, velocity, and timestamp information to achieve correspondence from frame to frame for a particular camera. Figure 7 shows how even when the CAMSHIFT algorithm does not lock onto the face, often it is still well centered on the occupant's body and hence would not effect the accuracy of the calculations. Figure 8, however, shows a case where the estimates could be thrown off significantly. Heuristics based on human behavior can help minimize these occurrences.

The second SPS uses the microphone array and a camera as complementary sensors. The microphone array used feeds into a Signalogic Sig32C-8 multichannel DSP board. The microphones were configured in 20x20 cm square arrangement and speech was sampled at 32 kHz. This provided us with approximately 19 samples of resolution assuming the speed of sound is 346 m/s. This rig provides a course means for measuring the 2D location of a speaker in a room. As a stand alone system, it was prone to picking up background noise from within the house and from outside. Being able to determine if the signal is actually a speech signal a priori greatly enhances its performance. Two methods were used to help determine the presence of voice. First, the short-term average magnitude of the audio signal was estimated using a recursive filter. It was then compared against a threshold value to determine if speech was present. This lightweight heuristic helped improve the false alarms considerably. With the addition of the face detection and recognition process, the system was further able to verify its measurements of occupancy.

The size of the room studied was 3.36 x 3.69 meters. The evidence grid for the Room Manager was broken down into two layers of resolution. First the room at a course level was divided into a 3x3 element where each square element was approximately 112x123cm. Each cell in the 3x3 grid was further decomposed into another 3x3 grid.

The data being sent to the RM from the SPSs is viewed as data from competitive sensors. While many techniques exist for fusing the data received from the SPSs, we opted first for simplicity. We took a simply histogram approach to populating the evidence grid. Figure 6 shows the contents of an evidence grid at the RM. The green areas in Figure 6 represents the regions of support for the hypothesis that an occupant is in a course grain grid element. The red area represents the most likely location of an occupant within that support region.



Incoming Sensor Data
Name: Scott
Sensor ID: Cam1
Cell ID: 14
Time: Wed Sep 12 10:11:29.92

Room Occupants
Scott: Cell 14
Irfan: Cell 55

Figure 6: *This represents the occupancy grid maintained by the Room Manager. The green region represents the region of support that someone is in that space. The red region represents a less certain area of support that estimates the occupant's location.*

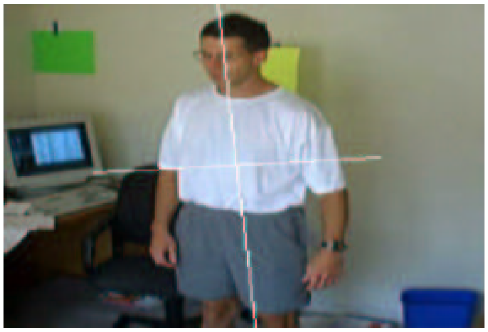


Figure 7: *This picture shows that even when the CAMSHIFT tracker fails to locate the face, it still locks onto the body and is well centered. These kind of errors have little affect on our method.*

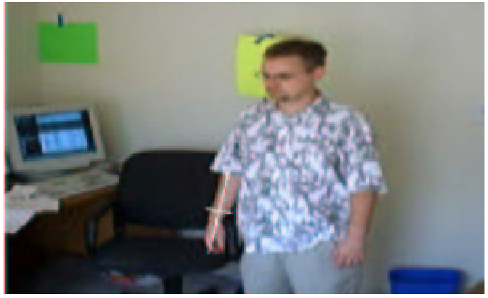
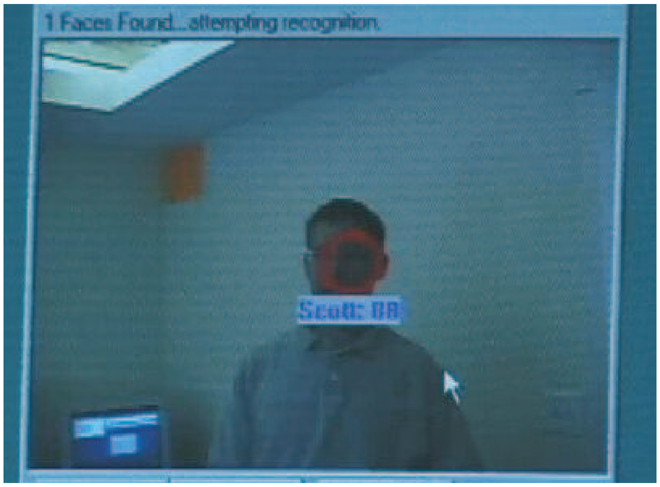


Figure 8: *This picture shows tracking errors that do impact our current approach. Further feature extraction and processing is required to resolve these errors.*

5. FUTURE WORK

The goal of this research is to develop a highly reliable means for tracking occupants moving about the house. A next step in our current work is to be able track the handoff transitions between room and hall way and hall way to other rooms as a resident moves about the house. This will require the development of a House Manager to maintain a presence on the state of each room. In addition, being able to generate hypothesis from recent past events and test their validity will help us further our journey toward homes that are truly aware.

Another soon to be incorporated capability for audio sensing is speaker identification. As it turns out, given the relatively low number of occupants in a house, preliminary analysis from our DSP group has shown that speaker identification can perform reasonably well in such an environment.



1 Faces Found...attempting recognition.

Figure 9: *This picture shows a successful recognition of an occupant. Successful face detection here can be used to question the kind of error shown in Figure 8.*

This capability will help fill the gaps when face recognition is not available and augment our visual analysis when both are available.

The prototype system presented here has to be incorporated with a toolkit that allows for easy instantiation of context sensitive actuators and sensors. It is for this reason, we are pursuing the addition of this prototypical system with Context Toolkit [9], which is serving as the primary context manager in the aware home project. To address the computational complexity of dealing with rich multimedia streams, we are also looking into parallel and distributed architecture for real-time processing [22].

Finally, we are also very interested in developing probabilistic methods to aid in the multimodal fusion. Towards this end, we are implementing Bayesian Inference mechanisms and belief propagation techniques that would reside at the process supervisor (PS) level.

6. ACKNOWLEDGMENTS

Thanks to the various members of the Aware Home project at Georgia Tech, especially Gregory Abowd and Kris Nagel, for they invaluable guidance with this work. Funding was provided by the Aware Home Research Initiative and by Georgia Tech's Broadband Institute.

7. REFERENCES

[1] S. Basu, M. Casy, W. Gardner, A. Azarbayejani, and A. Pentland. Vision-steered audio for interactive environments. In *Proceedings of IMAGE'COM 1996*, May 1996.

[2] A. Bobick, S. Intille, J. Davis, F. Baird, L. Campbell, Y. Ivanov, C. Pinhanez, A. Schütte, and A. Wilson. The KidsRoom: A perceptually-based interactive and immersive story environment. *"PRESENCE: Teleoperators and Virtual Environments"*, 8(4):367–391, August 1999.

[3] G. R. Bradski. Computer vision face tracking for use in a perceptual user interface. Intel Technology Journal Q2, 1998.

[4] R. R. Brooks and S. S. Iyengar. *Multi-Sensor Fusion*. Prentice Hall PTR, 1998.

[5] B. Brumitt, B. Meyers, J. Krumm, A. Kern, and S. Shafer. Easyliving: Technologies for intelligent environments. Proceedings of Handheld and Ubiquitous Computing, September 2000.

[6] J. Crowley and J. Bedrune. Integration and control of reactive visual processes. 1994 European Conference on Computer Vision, 1994.

[7] J. Crowley and F. Berard. Multi-modal tracking of faces for video communications. IEEE Conference on Computer Vision and Pattern Recognition, June 1997.

[8] J. L. Crowley and Y. Demazeau. Principles and techniques for sensor data fusion. *Signal Processing*, 32(1-2):5–27, May 1993.

[9] A. K. Dey, D. Salber, and G. D. Abowd. A conceptual framework and a toolkit for supporting the rapid prototyping of context-aware applications. *Human–Computer Interaction*, 16, 2001. To appear as anchor article in special issue on context-aware computing.

[10] R. Duraiswami, D. Zotkin, and L. Davis. Active speech source localization by a dual coarse-to-fine search. Proceedings of ICASSP2001, May 2001.

[11] H. F. Durrant-Whyte. Sensor models and multisensor integration. *International Journal of Robotics Research.*, 7(6):97–113, Dec 1988.

[12] A. Elfes. Occupancy grids: A stochastic spatial representation for active robot perception. *IEEE Computer*, 22(6):46–57, 1989.

[13] I. A. Essa. Ubiquitous sensing for smart and aware environments. *IEEE Personal Communications*, October 2000. Special Issue on Networking the Physical World.

[14] S. Goodridge and M. Kay. Multimedia sensor fusion for intelligent camera control. Proc. of 1996 IEEE/SICE/RSJ Int. Conf. on Multisensor Fusion and Integration for Intelligent Systems, December 1996.

[15] Y. Huang. *Real-time Acoustic Source Localization with Passive Microphone Arrays*. PhD thesis, Georgia Institute of Technology, February 2001.

[16] MIT. House_n project. http://architecture.mit.edu/house_n/web/, 2001.

[17] H. Moravec and M. Blackwell. Learning sensor models for evidence grids. Robotics Institute Research Review, 1992.

[18] E. D. Mynatt, J. Rowan, A. Jacobs, and S. Craighill. 2001 digital family portraits-supporting peace of mind for extended family members. Proceedings of CHI 2001, 2001.

[19] K. Nagel and G. ABowd. The family intercom: Developing a context-aware audio communication system. In Proceeding of Ubicomp 2001 International Conference, 2001. Technical Note.

[20] A. Pentland. Smart rooms. *Scientific American*, 274(4):68–76, April 1996.

[21] D. Rabinkin, R. Renomeron, A. Dahl, J. French, J. Flanagan, and M. Bianchi. A dsp implementation of source location using microphone arrays. *J. Acous. Soc. Am.*, 99(4):2503+, May 1996.

[22] U. Ramachandran, R. S. Nikhil, N. Harel, J. M. Rehg, and K. Knobe. Space-time memory: A parallel programming abstraction for interactive multimedia applications. In *Procceedings of 10th ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming*, 1999.

[23] S. Stillman and T. Tanawongsuwan. Tracking multiple people with multiple cameras. International Conference on Audio- and Video-based Biometric Person Authentication, March 1999.

[24] G. Tech. Aware home research initiative. http://www.cc.gatech.edu/fce/ahri/, 2001.

[25] J. Vermaak, M. Gangnet, A. Blake, and P. Patrick. Sequential monte carlo fusion of sound and vision for speaker tracking. Proceedings of ICCV 2001, July 2001.

[26] K. Yow and R. Cipolla. Feature-based human face detection. *Image and Vision Computing*, 15(9):713–735, 1997.