Active Camera Networks and Semantic Event Databases for Intelligent Environments

Mohan M. Trivedi, Ivana Mikic, Shailendra K. Bhonsle
Department of Electrical and Computer Engineering, University of California, San Diego

Abstract

In the future, intelligent rooms, with embedded multimodal sensory systems and semantic event databases, will support effective and efficient transactions of human activities and interactions. We are pursuing rigorous experimental investigations towards the development of such intelligent environments. In this paper we describe the overall system specification and general framework guiding our development. We also present details of the modules associated with the control and interpretation of video information acquired by a network of cameras and a novel semantic event database for characterization and recognition of activities. Evaluations of the system in a custom-built intelligent room are also presented.

Introduction

Intelligent environments provide numerous challenges in the machine perception area. It is strongly influencing recent research in the computer vision field [1]. Significant advances have been made in face recognition [2, 3], people tracking [4] and gesture recognition [5]. In audio analysis, much progress has been made in speaker and speech recognition [6] and source localization [7, 8]. There is a growing interest in multimodal systems, which integrate different modalities, such as audio and video [9]. One type of systems analyze gestures and spoken words for HCI applications to enable more natural "Put That There" type of interaction [10, 11]. Also, lip reading has been shown to improve speech recognition [12, 13]. An interesting system uses visual tracking information to drive the steering orientation of phased array microphones to emphasize audio input from the user that freely moves [14].

Our group is pursuing investigations for systematic development of Intelligent Environments where networks of cameras and microphone arrays serve as the sources of multimodal sensory information [8]. Figure 1 shows a conceptualization of an intelligent room with networks of cameras and microphone arrays, which are unobtrusively imbedded in the room. The research challenge is for the system to autonomously and robustly capture and maintain awareness of the objects and events in the space in a dynamic manner.

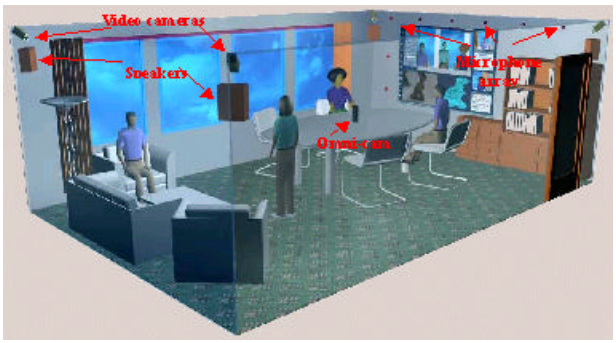


Figure 1. An Intelligent Room, with built-in infrastructure for efficient and effective transactions of human activities and interactions with participants from the same or remote physical spaces

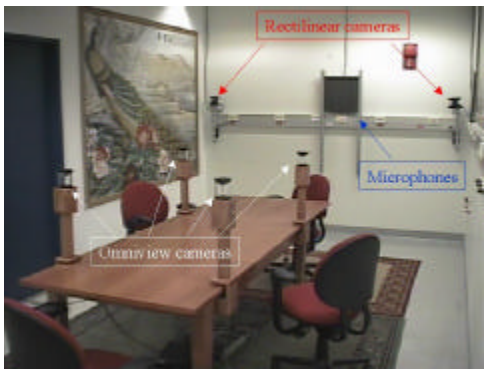


Figure 2. The AVIARY: Audio-Video Interactive Appliances, Rooms and sYstems. A testbed for experimental development and evaluation has four rectilinear cameras, four omnidirectional cameras and twelve microphones embedded in a room

The system design, development, and evaluation are accomplished in a multipurpose testbed called AVIARY for Audio-Video Interactive Appliances, Rooms and sYstems (Figure 2). These systems are required to *capture*, *process*, *transmit* and *display* audio-visual information in an integrated manner. Currently, a room is equipped with a network of four omnidirectional cameras, four pan/tilt/zoom rectilinear cameras and

twelve microphones. Robustness to environment is essential since it is not practical to dictate to the user a specific rigid environment. In addition, it is not unusual to expect the environment of the user to change, for example light getting turned on, blinds and curtains pulled up or the room furniture getting reconfigured. It is important that the systems still be able to carry out their task. To meet these requirements, systems need to be equipped with the following main capabilities:

- **Adaptive:** Systems must have the ability to adapt. This provides the systems the ability to deal with changes that take place in the environment.
- **Multimodal:** Systems must be equipped with proper information processing capabilities. This not only includes superior unimodal processing capability, but also well integrated multi-modal sensory information processing capabilities. This will result in more robust systems with broader range of capabilities
- **Active and Self-Calibrating:** Systems must have the ability direct is attention to "interesting" events and also to self calibrate.
- **Interactive:** Systems should be able to utilize semantic event databases to guide human-environment interactions. Semantic event databases store abstracted past states (events) of the intelligent environment and active sensory networks. These can be flexibly queried to guide human/automatic control of active sensory network and adaptation of processing and analysis algorithms.

Also, if mobility of the user is allowed, then handheld or head-mounted microphones are not appealing. Multiple microphones offer an attractive alternative. However, mobility comes at a price, and poses the following challenges:

- The received signal is distorted by the room acoustic properties, i.e. reverberations. This not only has ramifications on the quality of the audio signal, but also on any post-processing that maybe contemplated, e.g. speech recognition. Recovering the speech signal requires solving a difficult deconvolution problem. More importantly, the room acoustical properties are dependent on the user environment, can vary widely, and hard to account for a priori. Self-calibration is essential for the system to function in such diverse environments.
- The received signal is more susceptible to ambient noise and calls for clever spatial filtering techniques to enhance signal to noise ratio. In particular, the noise sources are not stationary (spatially and temporally) requiring systems to be adaptable.
- Vision system has to operate efficiently and robustly to conditions and events of the dynamic world. Novel approaches for real-time 3D modeling, real-time visualization and interaction, as well as real-time photorealistic rendering are required. The tracking of the speaker also has to be robust and fast.

Interactive systems developed in the AVIARY typically function either of the two modes: **(1) initialization or (2) active** mode**.** The system initialization refers to the situation where the system is learning its environment and carrying out the task of self-calibration in preparation for its run-time functions. The nature and abilities of the system during the system initialization dictates its robustness to environmental variables. During the active mode operation, the system implements useful functions such as acquiring and interpreting audio-visual signals thereby enabling effective interactions and providing a convenient user interface.

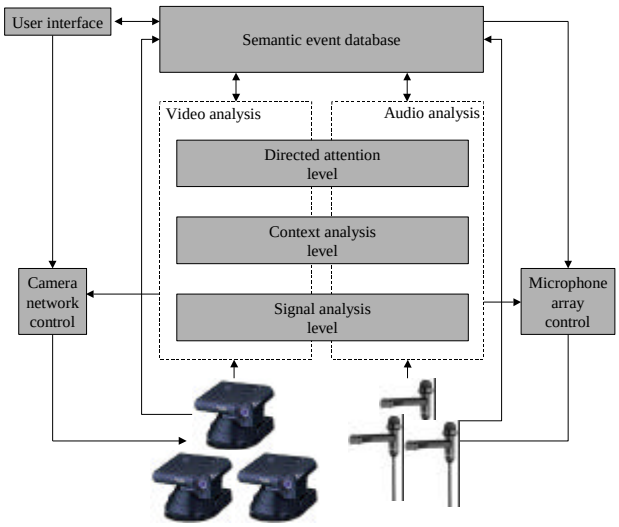


Figure 3. Active mode operational flowchart for human-environment interactions in the AVIARY.

Multi-level audio-video integration

The overall operational flow governing human-environment interactions in the AVIARY is summarized in Figure 3. As illustrated, the system is configured to fully utilize information provided by both audio and video sensory modalities. Integration of the audio-video information is accomplished at three levels:

- **Signal Anlysis Level:** At this level the system considers results of video segmentation, tracking

and audio source localization for robust localization and classification of objects and people

- **Context Analysis Level:** At this level the system performs face, voice, speech, gesture and pose recognition in an integrated manner. For example, face recognition and voice recognition algorithms would operate in concert for more reliable recognition. Other examples are integration of speech recognition and lip reading and "Put That There" type of multimodal interaction.

- **Directed Attention Level**: At the third level, analysis of one modality could *direct or focus the attention* of the part of the system that analyzes the other modality. One example would be in a teleconferencing setting with multiple people present. When one person starts to talk, their voice (or approximate location) would be recognized and attention of one or more PTZ cameras could be *directed* to that person's face. The video from one such camera would be transmitted alone, or combined with views from other cameras that have a view of the other participants. As the precise location of that person is determined by the video tracker, the phased array microphones could be steered toward them to enable accurate speech recognition. The system would also *focus* its attention on the "interesting" person by employing more sophisticated or expensive algorithms that are required for analysis of that person's actions, such as lip reading, facial expression or gesture recognition.

Omnidirectional camera network and audio processing modules

The omnidirectional vision sensor (ODVS) network embedded in the AVIARY has proven to be most useful for a number of tasks required in the system initialization and active mode operations. We have developed a multiple-, wide-baseline stereo system for accurate geometric modeling of the room [15]. We have also developed a visual modeling system, where accurate 3-D range information and color information are simultaneously extracted [16]. The ODVS network also allows us to track multiple people [17] and to dynamically generate the views associated with the movement of the persons in the room [18, 19]. The audio information is acquired using an array of twelve microphones and robust techniques for room acoustic modeling and speaker localization are being developed [8].

Rectilinear camera network and semantic event database

In the remainder of this paper, we will focus on the rectilinear camera network processing. We present our research related to a number of important and unique features of the "active" mode operation of the system. Specifically, we will be describe the following three important modules of the system.

- A robust and efficient human tracking module which utilizes a four-camera network.
- An active camera control for capturing frontal view of a person moving in the room. The module performs camera selection for best view and also automatic panning, tilting, and zooming for taking a close up image. This module provides inputs for the face detection and person recognizer modules.
- A semantic event database system which allows activity analysis (real-time as well as historical) using the results of the tracking module and a powerful language for characterizing complex activities as the spatio-temporal compositions of semantic activities.

Rectilinear camera network

We now concentrate on describing the analysis of video information, which allows realization of the above multilevel, multimodal sensory information integration.

Video Segmentation Module. The segmentation is based on background subtraction. First and second order statistics for background pixels are continuously updated. Due to the use of a forgetting factor, background model is adaptable to slow changes. Foreground pixels are segmented using the Neyman-Pearson test and grouped into blobs. Blob centroids for all cameras are computed and serve as input to the tracking algorithm. See [20] for details of the segmentation algorithm.

The Multi-Camera 3-D Tracking Module. The role of the tracker is to track multiple objects in 3D using segmentation results from different cameras with highly overlapping fields of view. The cameras are calibrated using Tsai's algorithm. For details of the algorithm, see [21]. The tracker is capable of tracking multiple objects simultaneously. It maintains a list of Kalman filters, one for each object in the scene. The real-time nature of the system requires the tracker to produce updated and predicted positions of each object for the current frame. Also, the availability of up-to-date prediction allows us to feed back the information to the segmentation

algorithm, which can increase its sensitivity in the areas where objects are expected to be present.

We evaluated the accuracy of the tracking algorithm on real data. Three different paths were measured and marked on the floor of a room (Figure 4). First set of sequences was recorded with three different persons walking each of the paths. For the second set, two people would walk at the same time on two different paths. Figure 5 shows ideal and measured paths for one experiment with two people. The tracks very accurately followed the ideal paths with maximum error in all experiments being around 200 mm and average error around 30 mm. That is a very good accuracy, especially if it's taken into account that the error in these experiments is influenced by calibration and segmentation errors and the errors in measuring the path, drawing it on the floor and walking on it.

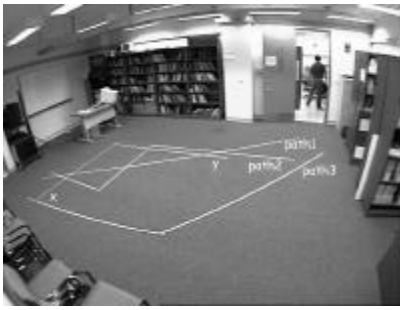


Figure 4. The world coordinate system and the measured paths

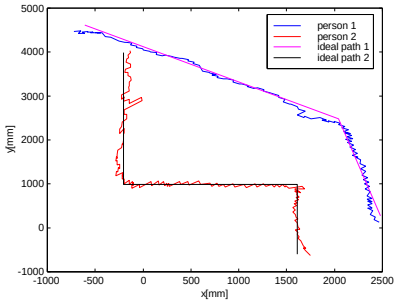
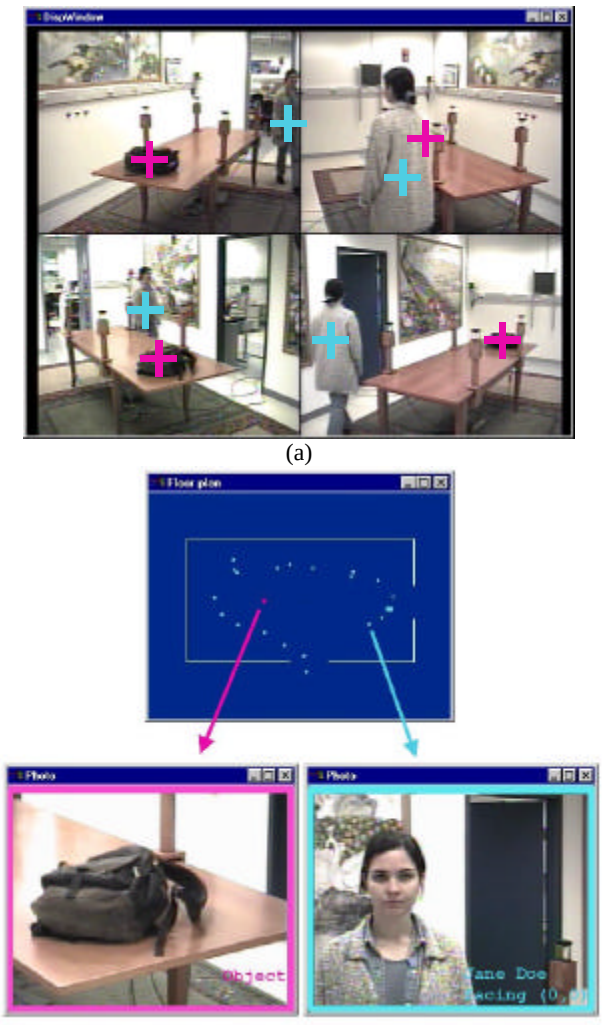


Figure 5. xy plots of tracks for two people in the scene

Active Camera Control Module. Four pan/tilt/zoom cameras (Canon VC-C3), controlled through an RS-232 link, are used in this system. Since the cameras are calibrated, it is trivial to compute pan and tilt parameters that bring an object at a known location to the center of view. Calibration of the zoom was done using the method described in [22], which enables us to find the zoom value that makes the size of the face in the image sufficient for recognition.

The current use of the PTZ capabilities of cameras is for taking snapshots of objects and of people's faces for face recognition and archiving. We also plan to use cameras for focusing of attention of the video-processing part of the system to "interesting" objects and locations.

As a new track appears in the room, it is classified as object or person based on its shape and size. If person is detected, the location of the head is estimated to be at the top fifth of the height. Best view camera is chosen to be the one for which the negative inner product of the viewing direction and average person's velocity in the last few frames is the largest. In other words, the camera that the person was walking toward in the recent frames is chosen to be the one that the person is most likely facing. If the track is classified as object, the snapshot of the whole object is taken and stored.



(a)

(b)

Figure 6. (a) example frame with the projections of 3D object locations onto the image planes (b) user has the view of the room floorplan with projections of tracks onto the floor plane. When a point belonging to a track is selected, the snapshot of the object/face is shown with the identification information.

Integrated Performance: *Person Tracking, Close-up Imaging, and Recognition*. Example of the system operation is given in Figure 6. The system interface shows the video from the four cameras with projections of the 3D tracks to the image planes shown as colored crosshairs. Also, a floorplan of the room is shown with projections of tracks on the floor plane. If user clicks on one of the tracks in this view, the snapshot of the face/object is shown with the recognition result.

Semantic Event/Activity Databases

In an intelligent room environment, there is the inherent notion of agents performing activities. Being able to retrieve *flexibly definable* past activities for analysis purposes is an important goal of such intelligent environments. Such analysis produces clusters *of usual behavior* that in turn can be used to detect unusual behavior in real time in such environments. For example, in a room usually people may walk around certain fixed paths to reach a computer, work on it, and then leave the region of the computer. An activity involving someone *climbing a table*, *reaching the computer*, *loitering* around the computer for some time, and then eventually *jumping off the table* may be an instance of an unusual activity. AVIARY architecture provides for a database framework for storing, flexibly defining, and retrieving activity information at high levels of semantic granularities.

At raw representational level, multiple sensors produce huge amount of sensory information. Such information can be analyzed to detect various activities, but the processing of this huge amount of data has prohibitive complexity. As a result, it is important to store and process only *meaningful* semantic information. One of the goals of such systems is to provide *abstractions* for dealing with meaningful semantic information. The multi-dimensional *feature sets* of objects is one such abstraction that is commonly used in the fields of multi-sensory, visual, and multimedia information systems. Objects of the system may be the observed mobile entities or spatial regions. Features are the basic domain-dependent semantic information associated with observed mobile objects or spatial regions. Finding other categories of high level semantic abstractions is an open and challenging field of research in processing of semantic information. The AVIARY database framework uses semantic abstractions of **events**, **activities**, and **behaviors** of associated environmental entities. An activity-of-interest is a complex spatio-temporal fusion of multi-sensory information over a large spatio-temporal extent. In our model we break down the complex process of defining a complex activity-of-interest into three simpler processes of defining events, defining activities through specification of spatio-temporal composition rules, and defining behaviors that are patterns of activities.

An event usually represents the state-transition of observed mobile objects or spatial regions. An event is a detectable atomic semantic unit, and the environment at any point in time is represented by a collection of occurred events. In the model presented here, we constrain an event to have bounded temporal extent. Such events represent complex spatio-temporal fusion of multi-sensory information in the intelligent environment scenario. The bounded extent defines the duration of state-transition that approaches zero, and includes the ***temporal uncertainty*** that is associated with either the ***detection*** or the ***definition*** of an event. An activity is a spatio-temporal composition of events. In this work, we focussed only on the temporal aspects of activity compositions. The database stores semantic events with their attributes (which include spatial parameters, agents, features etc.). Additionally and unlike commonly available database models, it also stores a specific type of temporal ordering information amongst events in the database. The query language provides facilities for the specification of activities through embedding activity composition rules in its language framework. The separation of semantic events (stored as the database instance) from semantic activities (embedded in queries to the database) achieves *flexibility* that is usually needed in semantic information processing in intelligent environments. Typically, only a few events need to be defined for detection using signal processing algorithms. Exponentially many activities based on user requirements can be composed on the fly using the query language. In the following we briefly discuss the data model, database system design, and provide an example of the use of the database in AVIARY.

Semiorder Data Model. Design of semantic event/activity databases depends on many factors. These factors include whether events represent state or state-transition of objects, whether spatial, temporal, or spatio-temporal aspect is emphasized, how spatial and temporal uncertainties in event occurrences are represented, whether the model of composition of activities from events is statistical or combinatorial, etc. For our modeling purposes, we defined events to represent state-transitions of observed objects of the system.

We focused on the temporal composition of activities from events, and an important consideration was to take into account the ***temporal uncertainties*** in event occurrences. Assuming that such uncertainties for a given set of events is bounded by a constant, we obtained a model of composition of semantic activities

from events. This model can suitably represent the ***concurrent occurrences of events*** in the presence of temporal uncertainties. The concurrency could be between events associated with ***multiple objects***, or it could also manifest itself between many events associated with a ***single agent***. Such concurrency in the presence of temporal uncertainties is *not modeled* by ubiquitous sequential temporal composition rules or by its simple extensions.

The combinatorial structure obtained above is a semiorder (a proper subclass of partial orders) based on the binary relation of *<precede,D>* such that *event x <precede, D> event y* if and only if the occurrence time of *y* is greater than the occurrence time of *x plus D*, where *D* is the fixed temporal uncertainty interval. Semiorders based temporal compositions represent a natural evolution of sequential composition rules. We designed a semiorder based data model and a corresponding query language that also embeds a semiorder pattern definition language. Many algorithmic and architectural issues associated with the design and implementation of this database are discussed in [23, 24].

System architecture for semantic event/activity processing. As mentioned above, semantic activities are complex spatio-temporal compositions of semantic events. The event/activity database stores events and their spatio-temporal inter-relations. A *transducer* detects these events using sensory information, spatial information stored in a surrogate spatial database, and feature data of detected objects. The event/activity database query language provides for high level semantic activity abstraction through its set of operations. These queries are executed against a database instance to detect activities. Also, there is provision for ***active queries*** that has query abstractions similar to those provided by the activity query language, but it executes in real-time and can notify different entities of the AVIARY architecture as soon as a high level activity is detected. Many components of such event/activity databases together with their interactions with the sensor processing elements and the event detection transducer are depicted in Figure 7.

The overall system architecture comprises of a visual signal processing subsystem described earlier in the paper, a transducer subsystem for detection of events, and the semiorder database subsystem for activity recognition. The database prototype is generic and can be used in other domains, for example in retrieval of appropriate visual information in visual information management. The semiorder database prototype is implemented in Java and works across different platforms. The schema definition, as well as events belonging to many *semiorder schemas*, is stored in a native object-oriented database. The schema is defined using a schema definition language, and is parsed to automatically generate needed schema and event classes.

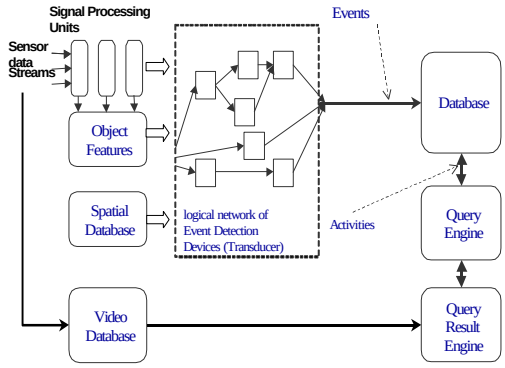


Figure 7. Detection and storage of semantic events and retrieval of semantic activities

The query environment consists of a query parser, a validator, an optimizer, an execution engine, and a query result visualization engine. The validator extracts the schema and validates the query against the schema definition. The current optimizer only decides *the set of data to be extracted from the native database*. It arranges *select* conditions in query expressions so that an optimal set of event data is extracted from the native database. The execution engine applies appropriate semiorder query language operators to this set of extracted data. Returned results always consist of sets of semiorders, and are navigated through and visualized using semiorder representations. We have a separate video database in our domain of application. The semiorder database query environment provides for retrieval of video sequences that *contain* retrieved activities.

Semantic Activity Recognition Results. A prototype of the semantic event/activity database was designed, developed, and used for detection of many complex *activities* in an intelligent room environment. A set of atomic events like startLeftTurn, endLeftTurn, enterRegion, exitRegion, Jumping, Merge (of objects), Split (of objects), Occur, Vanish, etc. were defined and corresponding *algorithms* were inserted in the event detection transducer. Many of these atomic events are defined only with respect to the *spatial structure* of the intelligent room. In our case, the spatial structure comprised of layers of spatial regions. Each such layer has disjoint spatial regions of interest. For example, we

divided the room into regions near doors, bookshelf, region of the computer systems etc.

Figure 8 depicts one of the results of a semantic query for detection of persons entering the region of the computer system after climbing the table. This is a simple example of a semantic query where different events form a totally ordered sequence. Many complex queries involving non-transitive *parallelism* between different events can be specified. The semantic event/activity database query mechanism is used to eventually analyze the intelligent environment to detect certain behaviors in real-time and/or to control the configuration of distributed clusters of sensors.

Figure 8. Demonstration of successful recognition of activities of a child using the semantic event database. A child *climbing* the table, *entering* the region of the computer, and *spending substantial length of time* in that region before *jumping off* the table are correctly recognized.

Concluding Remarks

Environments with rich and powerful suite of sensors allow more efficient means for transacting various activities among participants who may not be sharing the same physical space. We presented an overview of a systems oriented framework for design and specification of intelligent rooms. Two important functional blocks of the overall system are active camera network and semantic event database We presented the role, development and experiments associated with these blocks.

Acknowledgements

Our research is supported by the University of California Digital Media Innovation Program. Sponsors for the project include Sony Electronics, Compaq Computers Corporation and the state of California. We also appreciate the support of Canon Corporation (Japan) and valuable interactions with Trish Keaton of HRL. The authors gratefully acknowledge participation and contributions of Kim Ng, Kohsia Huang, Rick Capella, Nils Lassiter, Jonathon Vance, and Sadahiro Iwamoto to the overall AVIARY research.

References

[1] A. Pentland, "Looking at People: Sensing for Ubiquitous and Wearable Computing", *IEEE Trans. PAMI*, vol. 22, no. 1, January 2000, pp. 107-119

[2] A. Pentland, T. Choudhury, "Face Recognition for Smart Environments", *IEEE Computer*, vol. 33, no. 2, February 2000, pp. 50-55

[3] R. Chellappa, C. Wilson, S. Sirohev, "Human and Machine Recognition of Faces: A Survey", *Proc. IEEE*, vol. 83, no. 5, pp. 705-740, 1995

[4] D. Gavrila, "The Visual Analysis of Human Movement: A Survey", *Computer Vision and Image Understanding*, vol. 73, no. 1, January 1999, pp. 82-98

[5] V. Pavlovic, R. Sharma, T. Huang, "Visual Interpretation of Hand Gestures for Human-Computer Interaction: A Review", *IEEE Trans. PAMI*, vol. 19, no. 7, July 1997, pp. 677-695

[6] L. Rabiner, B. Juang, "Fundamentals of Speech Recognition", Englewood Cliffs, NJ: Prentice-Hall 1993

[7] M. S. Brandstein, "A Framework for speech source localization using sensor arrays", Ph.D. Thesis, Brown university, May 1995.

[8] M. Trivedi, B. Rao, K. Ng, "Camera Networks and Microphone Arrays for Video Conferencing Applications", *Proc. Multimedia Systems Conference*, September 1999

[9] R. Sharma, V. Pavlovic, T. Huang, "Toward Multimodal Human-Computer Interface", *Proc. IEEE*, vol. 86, no. 5, May 1998, pp. 853-869

[10] R. Bolt, "Put That There: Voice and Gesture at the Graphics Interface", *ACM Comput. Graph*, vol. 14, no. 3, 1980, pp. 262-270

[11] V. Pavlovic, G. Berry, T. Huang, "Integration of Audio/Visual Information for Use in Human-Computer Intelligent Interaction", *Proc. IEEE Int. Conf. Image Processing*, Santa Barbara, CA 1997

[12] C. Bregler, Y. Konig, "Eigenlips for robust speech recognition", *Proc. Int. Conf. Acoust. Speech Signal Processing*, 1994, pp. 669-672

[13] R. Kaucic, B. Dalton, A. Blake, "Real-time liptracking for audio-visual speech recognition applications", *Proc 4th European. Conf. Computer Vision*, 1996, pp. 376-387

[14] S. Basu, M. Casey, W. Gardner, A. Azarbayejani, A. Pentland, "Vision-Steered Audio for Interactive Environments", *Proc. IMAGE'COM '96*, Bordeaux, France, May 1996

[15] K. Ng, M. Trivedi, H. Ishiguro, "3D ranging and virtual view generation using omni-view cameras", *Proc. Multimedia Systems and Applications*, SPIE Vol. 3528, November 1998

[16] K. Ng, H. Ishiguro, M. Trivedi, T. Sogo, "Monitoring Dynamically Changing Environments by Ubiquitous Vision System", *IEEE Int. Workshop on Visual Surveillance*, June 1999

[17] T. Sogo, H. Ishiguro, M. Trivedi, "N-Ocular Stereo for Real-Time Human Tracking", Panoramic Vision: Sensors, theory, and Applications (R. Benosman and S.B. Kang, Editors), Springer Verlag, 2000

[18] K. Ng, H. Ishiguro, M. Trivedi, "Multiple Omni-Directional Vision Sensors (ODVS) Based Visual Modeling Approach", *IEEE Visualization*, October 1999

[19] K. Ng, H. Ishiguro, M. Trivedi, T. Sogo, "Human Tracking and Dynamic View Synthesis Using Network of Omni-Directional Vision Sensors", *Image and Vision Computing*, Special issue on Visual Surveillance (to appear)

[20] E. Sudderth, E. Hunter. K. Kreutz-Delgado, P. Kelly, R. Jain, "Adaptive Video Segmentation: Theory and Real-Time Implementation", *Proc. DARPA Image Understanding Workshop*, pp. 177-181, 1998

[21] I. Mikic, S. Santini, R. Jain, "Tracking Objects in 3D using Multiple Camera Views", *Proc. ACCV2000*, January 2000, pp. 234-239

[22] Collins and Tsin, "Calibration of an Outdoor Active Camera System", CVPR99, Fort Collins, CO, June 23-25, 1999, pp. 528-534.

[23] S. Bhonsle, "Semiorder Model for Temporal Composition of Activities from Events in Multi-Sensory Environments", Ph.D. dissertation, Dept. of Computer Science and Engineering, Univ. of Cal., San Diego, Winter 2000.

[24] S. Bhonsle, A. Gupta, S. Santini, M. Worring, R. Jain "Complex Visual Activity Recognition Using a Temporally Ordered Database", *Int. Conf. on Visual Information Management*, June 1999