Title

Dynamic context capture and distributed video arrays for intelligent spaces

Journal

IEEE Transactions on Systems Man and Cybernetics Part A-Systems and Humans, 35(1)

ISSN

1083-4427

Authors

Trivedi, Mohan Manubhai
Huang, K S
Mikic, I

Publication Date

2005

DOI

10.1109/TSMCA.2004.838480

Peer reviewed

Dynamic Context Capture and Distributed Video Arrays for Intelligent Spaces

Mohan Manubhai Trivedi, *Senior Member, IEEE*, Kohsia Samuel Huang, *Member, IEEE*, and Ivana Mikić, *Member, IEEE*

***Abstract*—Intelligent environments can be viewed as systems where humans and machines (rooms) collaborate. Intelligent (or smart) environments need to extract and maintain an awareness of a wide range of events and human activities occurring in these spaces. This requirement is crucial for supporting efficient and effective interactions among humans as well as humans and intelligent spaces. Visual information plays an important role for developing accurate and useful representation of the static and dynamic states of an intelligent environment. Accurate and efficient capture, analysis, and summarization of the dynamic context requires the vision system to work at multiple levels of semantic abstractions in a robust manner. In this paper, we present details of a long-term and ongoing research project, where indoor intelligent spaces endowed with a range of useful functionalities are designed, built, and systematically evaluated. Some of the key functionalities include: intruder detection; multiple person tracking; body pose and posture analysis; person identification; human body modeling and movement analysis; and for integrated systems for intelligent meeting rooms, teleconferencing, or performance spaces. The paper includes an overall system architecture to support design and development of intelligent environments. Details of panoramic (omnidirectional) video camera arrays, calibration, video stream synchronization, and real-time capture/processing are discussed. Modules for multicamera-based multiperson tracking, event detection and event based servoing for selective attention, voxelization, streaming face recognition, are also discussed. The paper includes experimental studies to systematically evaluate performance of individual video analysis modules as well as to evaluate basic feasibility of an integrated system for dynamic context capture and event based servoing, and semantic information summarization.**

***Index Terms*—Active vision, activity summarization, ambient intelligence, body modeling, event analysis, face detection/recognition, human–machine interfaces, multicamera systems, person tracking, real-time vision, smart rooms/spaces.**

I. Introduction

INTELLIGENT environments are indeed complex systems, where humans and machines (i.e., rooms) collaborate to accomplish a task. From such a perspective, intelligent environments can also be considered as a novel human–machine interface. The overall goal of intelligent environment research is to design and develop integrated sensor-based systems that allow natural and efficient mechanisms for human–computer interactions in places where humans work, learn, relax, and play. There is a growing interest in developing intelligent or smart spaces, and, like most new areas of research, there may not be a well-accepted definition for such terms. One possibility to address this issue could be to specify requirements, which a physical space needs to possess in order to be called intelligent. We consider the following four requirements in developing intelligent environments.

1) Intelligent spaces are designed for humans, and they should facilitate normal human activities taking place in these spaces.
2) Intelligent spaces should automatically capture and dynamically maintain an awareness of the events and activities taking place in these spaces.
3) Intelligent spaces should be responsive to specific events and triggers.
4) Intelligent spaces should be robust and adaptive to various dynamic changes.

Such spaces need not be limited to rooms in buildings, but extend to outdoor environments and any other spaces that humans occupy such as a performance on a stage, or an automobile on a highway. Design of such spaces is indeed a rather ambitious effort, especially when one considers the real-world challenges of providing real-time, reliable, and robust performance over the wide range of events and activities, which can occur in these spaces.

Novel multimodal sensory systems are required to realize useful intelligent spaces. Arrays of cameras and microphones distributed over the spatial (physically contiguous or otherwise) extent of these spaces will be at the front end of capturing the audio-visual signals associated with various static and dynamic features of the space and events. The intelligent environments will have to quickly transform the signal-level abstraction into higher level semantic interpretation of the events and activities.

The spaces are monitored by multiple audio and video sensors, which can be unobtrusively embedded in the infrastructure. To avoid intrusion on the normal human activities in the space, all sensors, processors, and communication devices should remain invisible in the infrastructure. The system should also support natural and flexible interactions among the participants without specialized or encumbering devices.

Manuscript received November 5, 2003; revised April 1, 2004 and June 18, 2004. This work was supported in part by a number of sponsors, including the California Digital Media Initiative and the University of California Discovery Grants projects, Technical Support Working Group, U.S. Department of Defense, Sony Electronics Corporation, Compaq Computers, and the DaimlerChrysler Corporation. This paper was recommended by Guest Editor G. L. Foresti.

M. M. Trivedi and K. S. Huang are with Computer Vision and Robotics Research Laboratory, University of California, San Diego, CA 92093-0434 USA (e-mail: mtrivedi@ucsd.edu; khuang@ucsd.edu).

I. Mikic is with Vala Sciences, San Diego, CA 92121 USA (e-mail: imikic@valasciences.com).

Digital Object Identifier 10.1109/TSMCA.2004.838480

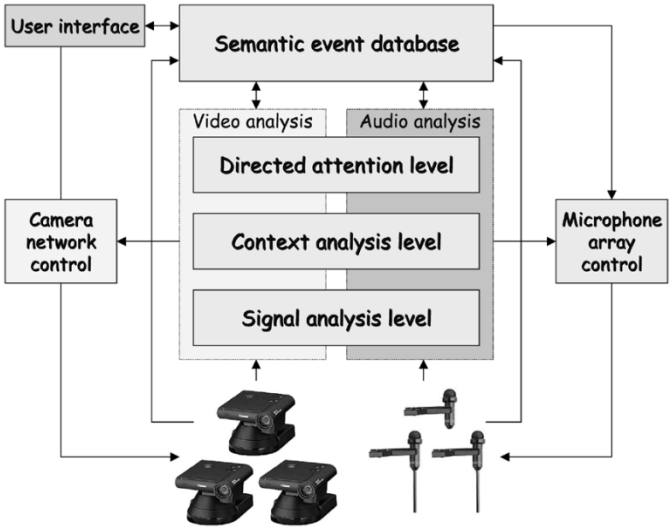


Fig. 1. Multilevel hierarchy of computational tasks associated with an intelligent environment. The system captures multimodal sensory signals and transforms it to a higher semantic level of information in order to facilitate human activities taking place in these spaces.

In an intelligent environment, multiple video cameras and microphones may be embedded in walls and furniture. Video and audio signals are analyzed in real time for a wide range of low-level tasks including: person identification, localization, and tracking; and gesture and voice recognition. Combining the analysis tasks with human face and body synthesis enables efficient interactions with remote observers, effectively merging disjoint spaces into a single intelligent environment. Fig. 1 shows the overall system conceptualization, functional blocks, and information flow associated with an Intelligent Environment. Multimodal sensory arrays capture signals from audio and video domains. These signals are represented in a time-synchronized manner using appropriate basis functions. Classification algorithms allow extraction of higher level semantic information from the signals. Such interpretation along with task specifications generates control signals for the sensory arrays for acquiring the next set of signals, from only an *attention zone* at a selected spatial-temporal resolution. Successful operation of the intelligent environment requires it to operate as a finely tuned system, where information resides at multiple levels of abstractions. Key levels to consider are.

1) Signal: This is the lowest level of abstraction where signals from multi modal sensors are captured and represented digitally in the forms of pixels, optical flow, pitch, or cepstrum.
2) Object: This is a pattern defined in the spatial domain. We focus on objects, which are defined using video sensory modality. Examples of such objects would be a person or face.
3) Event: This is a pattern defined in the spatial-temporal domain. We consider events using both audio and video modalities. Examples of events can be a person entering/leaving a room, or a person speaking.
4) Activity: This is a complex (or compound) pattern of events. We consider activities using both audio and video modalities. Examples of an activity can be people having a meeting or a person dancing in a room.
5) Context: This is considered to be a specification of the state of an intelligent environment. It is defined using prior knowledge of the environment and tasks. Events detected from sensory information would cause changes in the state of the system.

Recent research on intelligent environments provides numerous new challenges in the fields of machine perception. In computer vision [1], distinct progress in face detection and recognition [2]–[5], people tracking [6], [7], and gesture recognition [8], [9] has been made in the last decade. For audio, much progress has been made in speaker and speech recognition [10] and source localization [11], [12]. Integrated sensory modalities of audio and video [13]–[18] are also being seriously considered recently. One type of system that recognizes gesture and spoken words made possible a more natural "put that there" type of interaction between humans and computers [19]. We are currently embedding distributed video networks in rooms, laboratories, museums, and even outdoor public spaces, in support of experimental research in this domain [20]. This involves the development of new frameworks, architectures, and algorithms for audio and video processing, as well as for the control of various functions associated with proper execution of a transaction within such intelligent spaces. These test beds are also helping to identify novel applications of such systems as distance learning, teleconferencing, entertainment, and smart homes.

In this paper, we present a framework for efficiently analyzing human activities in the environment, using networks of static and active cameras. Information will be extracted at multiple levels of detail, depending on the importance and complexity of activities suspected to be taking place at different locations and time intervals. The environment will be constantly monitored at a low resolution, enabling the system to detect certain activities and to estimate the likelihood that other more complex activities are taking place at specific locations and times. If such an activity were suspected, to enable its accurate perception, a higher resolution image acquisition and more sophisticated analysis algorithms would be employed. Current systems focus on analyzing data at a fixed resolution, in some cases, monitoring a large space with a single camera and in others covering a small area with many cameras. We believe that the middle ground has not been sufficiently explored and that combining the coverage and robustness of low-resolution analysis with the power of high-resolution analysis will result in robust and efficient systems that will be capable of extracting high quality, relevant information from the environment.

The paper includes details of this multiresolution computational framework to help design the distributed video arrays (DIVA) for intelligent environments. We also describe the infrastructure and experimental testbeds of utility in design and evaluation of indoor intelligent spaces. We will focus on real-time tracking of single or multiple people and on coordination of multiple cameras for capturing visual information on wide areas as well as selected areas for activity analysis and person identification. Finally, a detailed design and experiments conducted in an intelligent meeting room are presented.

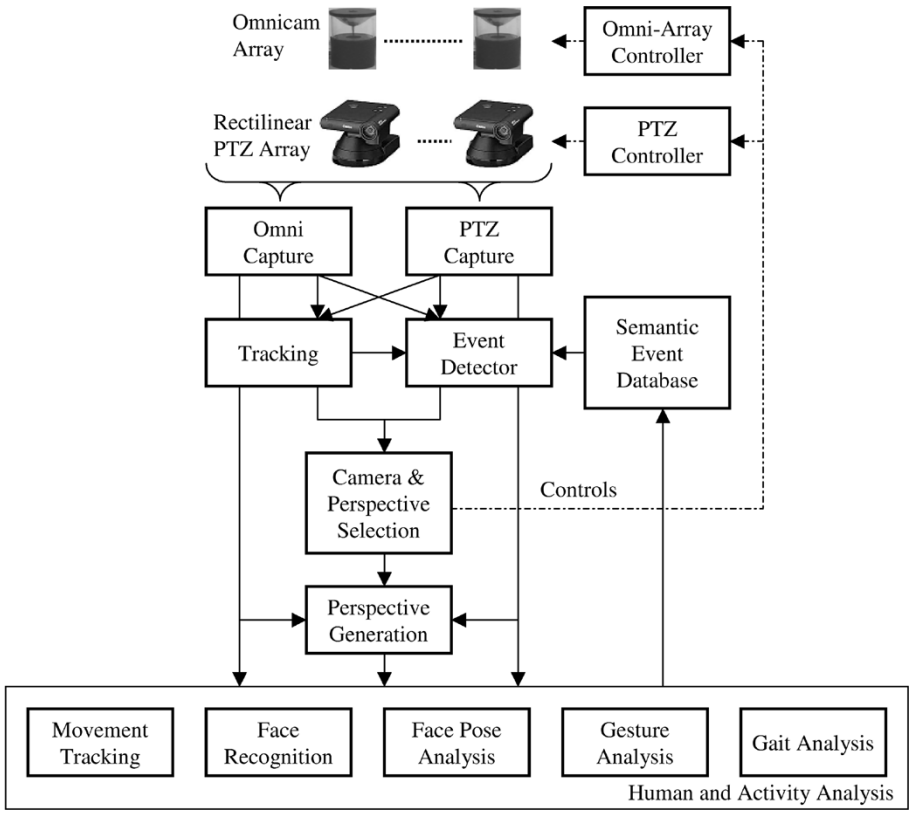


Fig. 2. DIVA for tracking, human identification, and activity analysis.

II. DIVA for Intelligent Environments

DIVA is an intelligent environment that is able to detect the presence of people, track their movements, recognize them, and understand their actions, as shown in Fig. 2. For recognition of complex actions, high-resolution images of the human body (or even body parts) are necessary. To allow unconstrained human movement in a large surveillance area, the system should therefore be able to acquire such high-resolution video anywhere in the environment. Some computer vision groups have equipped their laboratories with large numbers of static cameras [21], [22] with the idea of obtaining very detailed information about the monitored space. However, on the other hand, for the purpose of maintaining awareness of the presence and activities of people, the system does not need detailed information all the time and everywhere in the environment, but only at specific intervals or locations when/where something potentially interesting is happening. At other times, much less detail is sufficient. Detecting a person's presence or recognizing whether they are sitting or standing requires less detailed information than estimating the direction that the person is pointing his/her finger to. Based upon these observations, we propose a system that continuously monitors the environment at low resolution, which detects only the presence and location of people and their dimension. More detailed image acquisition and analysis would be triggered when a potentially interesting event or activity is suspected to be taking place. We will term those potentially interesting events as the focuses of attention of the system. Equipped with a few static wide-angle view cameras, the low resolution but large area monitoring of the environment can be achieved. With a small number of active pan/tilt/zoom (PTZ) cameras, multiple simultaneous focuses of attention can be maintained. Depending on the activity to be analyzed, the active camera can focus on various levels of details of people, from the whole body to face or hand gestures. Using this approach, robust monitoring of the entire environment with multiple resolutions can be achieved with fewer cameras and computational resources.

In this section, we discuss the development of such DIVA, which support a wide range of tasks of the intelligent environments. Key features of these smart video arrays are:

1) the ability to derive semantic information at multiple levels of abstraction;
2) the ability to be attentive to specific events and activities;
3) the ability to actively shift the focus of attention at different semantic resolutions;
4) the ability to apply different types of camera arrays to provide multiple signal-level resolutions.

To develop such a multilevel approach, problems of camera placement and control, as well as the designing of image-analysis algorithms have to be addressed. Good camera placement will provide efficient coverage. The control problem involves developing the system that will acquire data from certain locations/time intervals in the environment and employ appropriate analysis algorithms at the level of detail needed to maintain awareness of the people and their activities. This may often involve maintaining multiple simultaneous focuses of attention.

Algorithms that track people in three-dimensions (3-D) at multiple resolutions are essential parts of the proposed system.

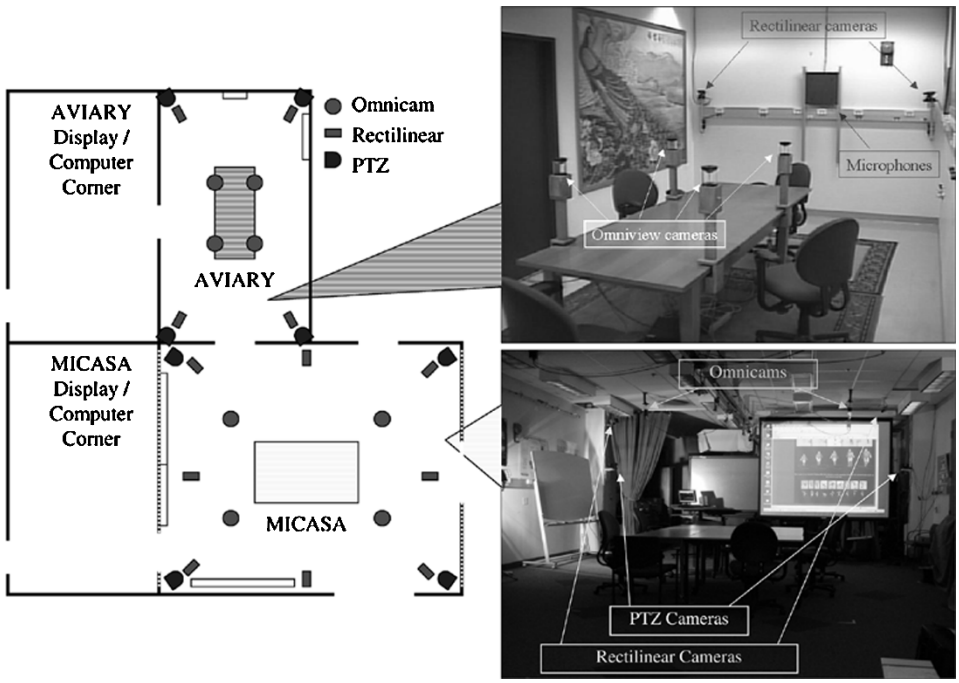


Fig. 3. Floor plan and camera network configurations of our intelligent space complex. These rooms are built for experimental development and evaluation of the intelligent room systems utilizing 12 rectilinear cameras, eight omnidirectional cameras, eight PTZ cameras, and eight microphones, which are embedded in the room.

At a low-resolution level, locations of all people in the environment will be continuously monitored. A more sophisticated algorithm is needed to extract more detailed body posture and motion estimates. The algorithm should be able to extract multiple levels of detail depending on the quality of the available data and the level of detail currently requested by the system.

Camera videos are first captured and processed for low-level visual cues, such as histograms, colors, edges, and object segmentations. The challenges at this level include: robustness to illumination, background, and perspective variations.

On the next level of abstraction, tracking plays an important role in event analysis. It derives the current position and geometry of people as well as the histories and predictions of their trajectories. With the semantic database, which defines prior knowledge of the environment and activities, events can be detected from the tracking information, e.g., one person enters the room and sits beside a table. The activity analyzer and the semantic database could be implemented by a rule base or a Bayesian net [23]. The challenges at this level include: the speed, accuracy, and robustness of the tracker, as well as the scalability of the semantic database, which allows incremental updating when new events are detected.

The events trigger the attention of a camera array to derive higher semantic information. Using tracking information, a suitable camera is chosen to capture a perspective that covers the event at a desired resolution, e.g., perspective on a person with an omnicam for posture and around the head area with a PTZ camera for person identification. For this purpose, necessary processing modules, such as face detection and recognition, should be deployed. The challenges at this level include: accuracy, speed, and robustness of the view generation and recognition modules. The derived semantic information at multiple levels can also be fed back to update the semantic database.

This architecture of multilevel abstraction can be further generalized to include many other applications such as object recognition, facial expression recognition, 3-D human-body modeling and tracking [24], and behavior estimation and prediction [25].

The DIVA system architecture developed for the intelligent environments described in this paper can be viewed as a smart or active camera network, where various cameras are actively controlled to support a wide range of functionalities. It is recognized that the proper operation of the overall system depends on the success of selecting the proper parameters for video capture, perspective selection, feature extraction, object/event detection, tracking, storage/archiving, and interactions with humans. It is also important to emphasize that video streams which are primary inputs to DIVA at the raw pixel levels are prohibitively large in size. Success of the system will depend on the ability to eliminate redundancy, to transform raw data into higher semantic levels, and to be very selective in acquiring new video data only when needed and also from a specific region of interest and at the appropriate resolution. These observations help in designing video arrays which get turned on only when needed and vision algorithms which extract the context specific cues to support proper operation of the overall system. In this paper, we have focused only on the intelligent environments which physically continuous entities like conference room and performance space. However, the active vision concepts used in the DIVA architecture allow them to be effective in monitoring and surveillance applications of very large distributed spaces, such as highways and open public spaces [20].

III. Intelligent Environments: System Infrastructure and Experimental Testbeds

Systematic development of intelligent environments, where networks of cameras and microphone arrays serve as the sources of multimodal sensory information, is indeed a system-oriented experimental research effort. In this section, we present the overall infrastructure and some novel experimental testbeds designed to support design and evaluation of computational modules. We also discuss the experimental system architecture of our intelligent space complex.

A. Intelligent Environment Research Complex

The intelligent environment research complex at the Computer Vision and Robotics Research (CVRR) Laboratory, University of California, San Diego, is shown in Fig. 3. It includes two separate but connected rooms appropriately instrumented and suitable for a wide range of experimental research. The first one is audio video interactive appliances, rooms, and systems (AVIARY), which was designed to be a small conference room. The second space is multimodal interfaces, and context aware spaces (MICASA), which was designed to be a classroom or a performance chamber. We present a brief overview of these testbeds below.

The audio-video sensory suite installed in the AVIARY room includes a network of four omnidirectional cameras, four PTZ and four static rectilinear cameras, and eight microphones. The four omnicameras (ODVSs), are near the corner of a meeting table, covering the entire room from inside out. ODVS is a catadioptric camera with a hyperboloidal mirror to cover a downward hemispherical field of view [26]. The omnidirectional video can be unwarped into either a panoramic video or a PTZ rectilinear video by nonlinear transformations [6]. The four static rectilinear cameras are located at the upper four corners of the room, each covering the entire room from outside in. This directional difference matters with tracking performance as will be mentioned later. Also, four PTZ rectilinear cameras are installed at the four corners about 1.4 m above ground. They capture events with higher resolutions than the static cameras but narrow field of view. Two microphone arrays, each with four microphones, are, respectively, installed on the wall and the ceiling to pick up the speech activities in the meeting. A white board is sitting at the upper right corner of the room as shown in Fig. 3. One computer resource is allotted to tracking, which takes either the four static omnicam videos or the four static rectilinear videos. Another computer is used to analyze audio and visual events within the room. The third computer is used to archive the audio and video streams for later retrieval. AVIARY is used to develop and evaluate systems that capture, process, transmit, and display audio-visual information in an integrated manner. The audio and video modalities provide valuable redundancy and complementary functionality. These two modalities are also the most natural ways for humans to sense and interpret their environments, and interfaces of these two modalities can be very natural and effortless for the users. Robustness to the environment is another essential requirement 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, lights getting turned on, or the room furniture getting reconfigured. It is important that the systems still can carry out their task.

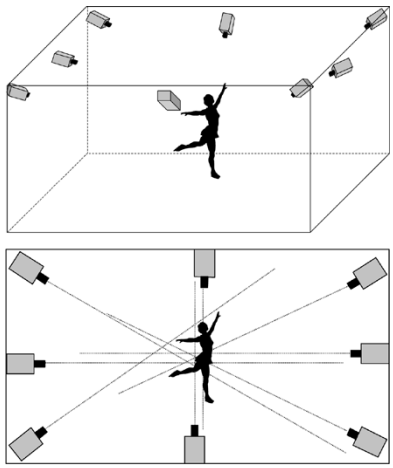


Fig. 4. MICASA static rectilinear camera array placement.

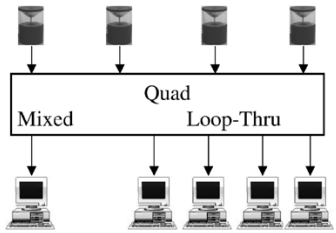


Fig. 5. Architecture for synchronous video capturing using quad.

MICASA is two times larger than AVIARY. The omnicam array is installed on the ceiling to cover the entire space. The PTZ rectilinear camera array is installed similar to AVIARY. However, there are eight static rectilinear cameras installed on top of the room, as shown in Fig. 4. The four cameras at the corners have larger field of view to cover the entire room and can serve as the tracking camera array. The other four have smaller coverage for a little better detail. All eight overlap each other by approximately a $2 \times 3 \times 2.5$-m volume. Within this volume, voxel reconstruction of human objects can be performed by shape-from-silhouette. The pairs of cameras that face each other are placed with offset, since the two cameras that directly face each other collect redundant 2-D silhouette information of the object. The camera videos are captured frame-by-frame synchronously. For the computational resources, currently, one PC is dedicated to tracking with the omnicam array. More PC would be favorable to increase the resolution of tracking. Six other PCs are allotted to voxel reconstruction with six of the eight static rectilinear cameras. Currently, no microphone arrays are installed in MICASA. A projector presentation board is sitting at the left side and a white board is sitting at the lower left-hand side of the room for classroom setup, as shown in Fig. 3.

B. Camera Calibration

Camera calibration affects the tracking and voxel reconstruction accuracy. The static rectilinear cameras are calibrated automatically using Tsai's algorithm with respect to a unique world-

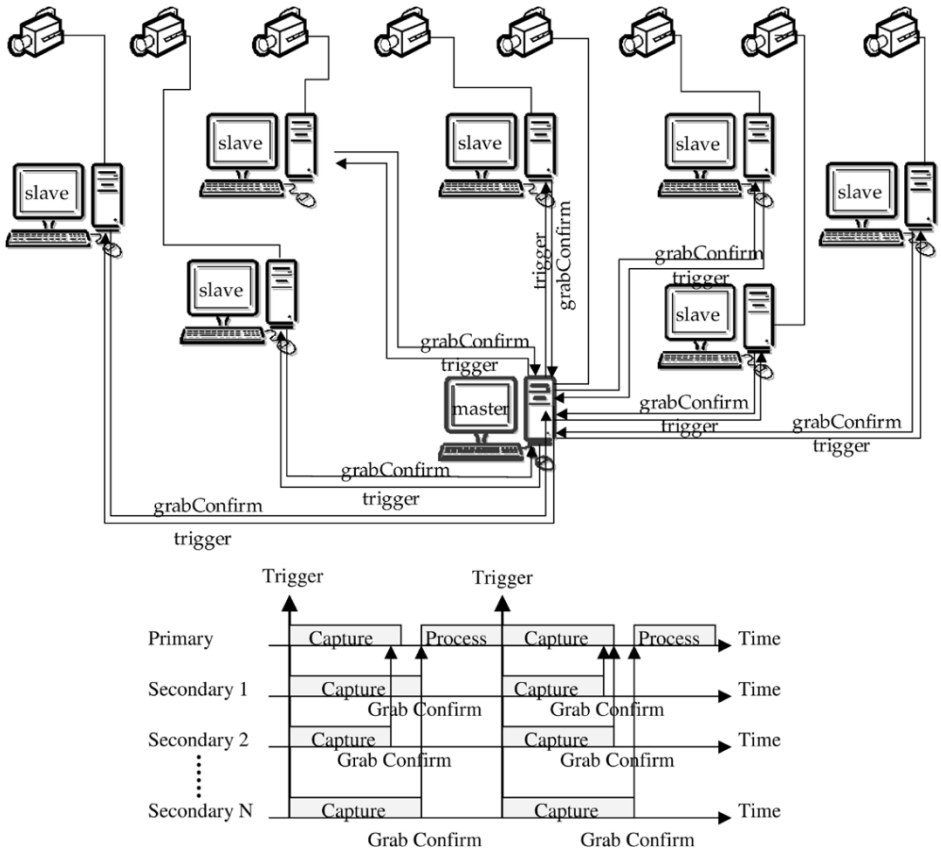


Fig. 6. The primary-secondary architecture for full-frame synchronous capture of multiple video streams.

coordinate system [27]. The calibration is carried out in advance and parameters are stored in the computers.

The calibration of ODVS is carried out manually. We collect a set of calibration points with their coordinate values in a world frame with the origin at one corner of the room. The world coordinates of the ODVS optical center are also taken. If the ODVS is sitting upright, then the absolute azimuth orientation of the ODVS can be estimated by rotating the relative direction of the calibration points in the omnidirectional image around the center of the image to match the azimuth directions of the calibration points with respect to the optical center of the ODVS in the world coordinate frame. The way to see whether the ODVS is sitting upright is by checking whether a set of markers at the same height as the optical center of the ODVS is on a concentric circle in the omnidirectional image that corresponds to the horizontal level, or whether they align on a row in the unwarped panorama that corresponds to the horizontal level. If the ODVS is tilted, then the absolute orientations of the camera need to be estimated analytically by relating the world coordinates and the camera coordinates with the mirror optics. However, an approximate approach may be taken if the tilting is very small. From the horizontal markers mentioned previously, we can tweak around the center of the omnidirectional image by several pixels to make the horizontal markers align with the horizontal row of the unwarped panorama. This approximation is used to improve the tracking accuracy in our experiments.

C. Synchronized Video Capture in DIVA

Arrays of cameras are included in the DIVA system to capture visual cues in the overlapped zone in a synchronized manner. For the omnicam, static rectilinear, and PTZ rectilinear camera arrays, three approaches of frame synchronization on video capturing may be taken. The first one is to use quad video multiplexers to combine four videos into one to be captured by the computer image grabber, as shown in Fig. 5. This, by nature, guarantees synchronized capture of the four camera videos. However, image resolution of each camera is reduced to one fourth. In larger space such as MICASA and applications that require fine details, this may be unsatisfactory. As the second approach, each video is captured by a computer in full frame and synchronized by time stamps. This approach allows preprocessing to derive some higher-level visual cues of each full-frame video before sending them through network to a server for integrated analysis. However, since the clock cycles of the video frames are not genlocked between the cameras, the time-stamp synchronization is only approximate with errors as much as 17 ms. For real-time voxelization of moving subjects like waving arms, millisecond jitters of capture timing can cause large misalignment between each camera array frames and deteriorate voxel reconstruction. Also, network traffic jam would reduce the frame rate of real-time human tracking.

The third approach is to synchronize the image grabbing by hardware devices, as shown in Fig. 6. This way guarantees full-

Fig. 7. Workstations that perform synchronous grab from multiple cameras.

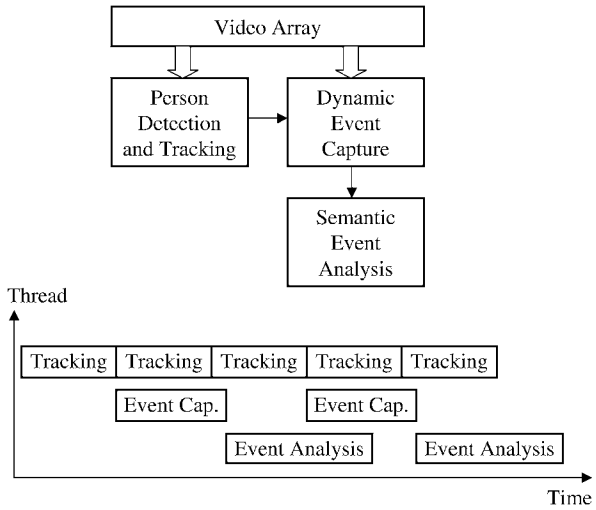


Fig. 8. System processes and multithread synchronization for active event capturing.

frame capturing for high-resolution demands as well as capture timing. In our systems, each camera is connected to a PC with a Matrox Meteor II frame grabber. To ensure synchronous grabbing, one PC is designated as the primary and the others as secondary. The primary sends a trigger signal to the secondary frame grabbers, which grab a frame at the rising edge of the trigger pulse. The trigger pulse is boosted and distributed to the secondary machines using a high-speed CMOS 74HC244 octal inverting tristate buffer. Each output of the octal buffer is connected to an RG58 cable, and can then be attached to a Meteor II input cable for external triggering. Multiple boosters can be cascaded if more than eight videos are needed. Additionally, the signal can be converted to a RS-232 computer serial port signal to allow for alternative triggering methods, e.g., synchronization of frame grabbing from firewire cameras via serial port. The set of workstations used is shown in Fig. 7.

D. Active Control for Event Capture in DIVA

The DIVA system is designed to capture the interested objects and events in the sensor array coverage, as shown in Fig. 8. Person detection and tracking is carried out on the static video arrays. Multiple baseline stereo on the synchronized static video arrays measures the locations of people on each frame and tracking filters smooth measurement noises and predicts the trajectories of people. When the trajectory is

TABLE I
SUMMARY OF THE INTELLIGENT COMPLEX SETUP

	AVIARY	MICASA
Size	6.7m × 3.3m ×2.9m	6.7m × 6.6m ×2.9m
Video Array	4 Omnicams 4 Rectilinear 4 PTZ	4 Omnicams 8 Rectilinear 4 PTZ
Video Synchronization	Quad synchronized	Quad or hardware synchronized
Audio Array	2 microphone arrays	None
Processors	1 for tracking 1 for event analysis 1 for video & audio archiving	1 for tracking 6 for voxelization and event analysis

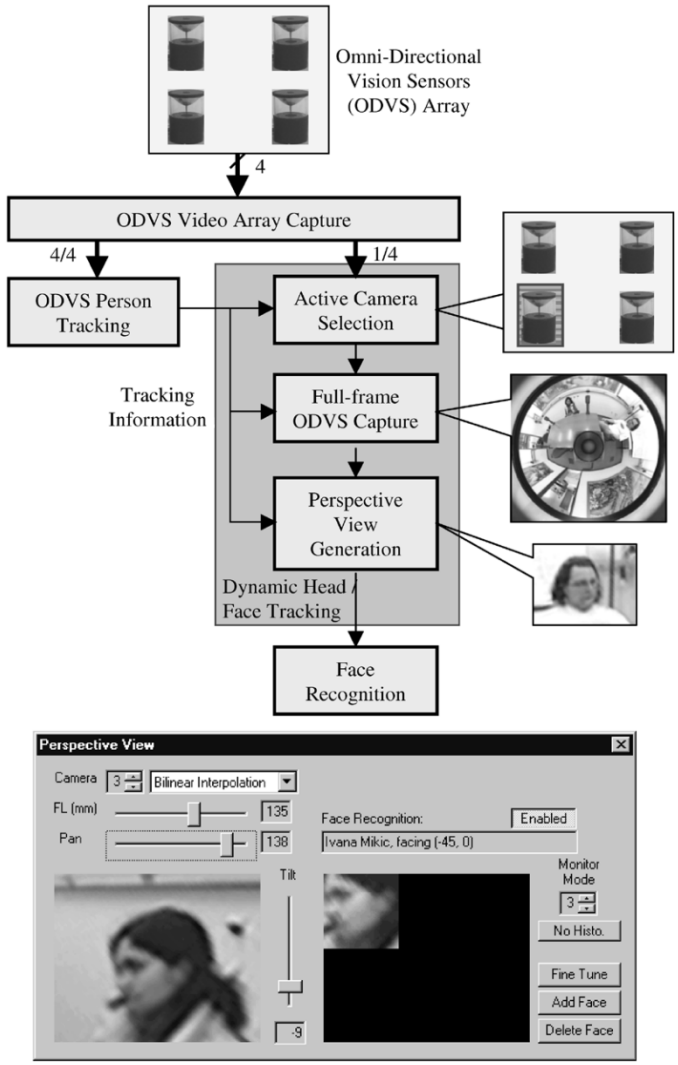


Fig. 9. Functional blocks of the NOVA intelligent room system. The lower window shows head tracking by ODVS perspective view generation and face recognition.

available, low-level events, such as a person entering the room or a person sitting down, triggers system attention. The system then captures more details of the event by driving a dynamic camera to it, and higher level analysis and interpretation of the event is computed. The processes are implemented in C++ with multithreaded programming, and the thread synchronization is

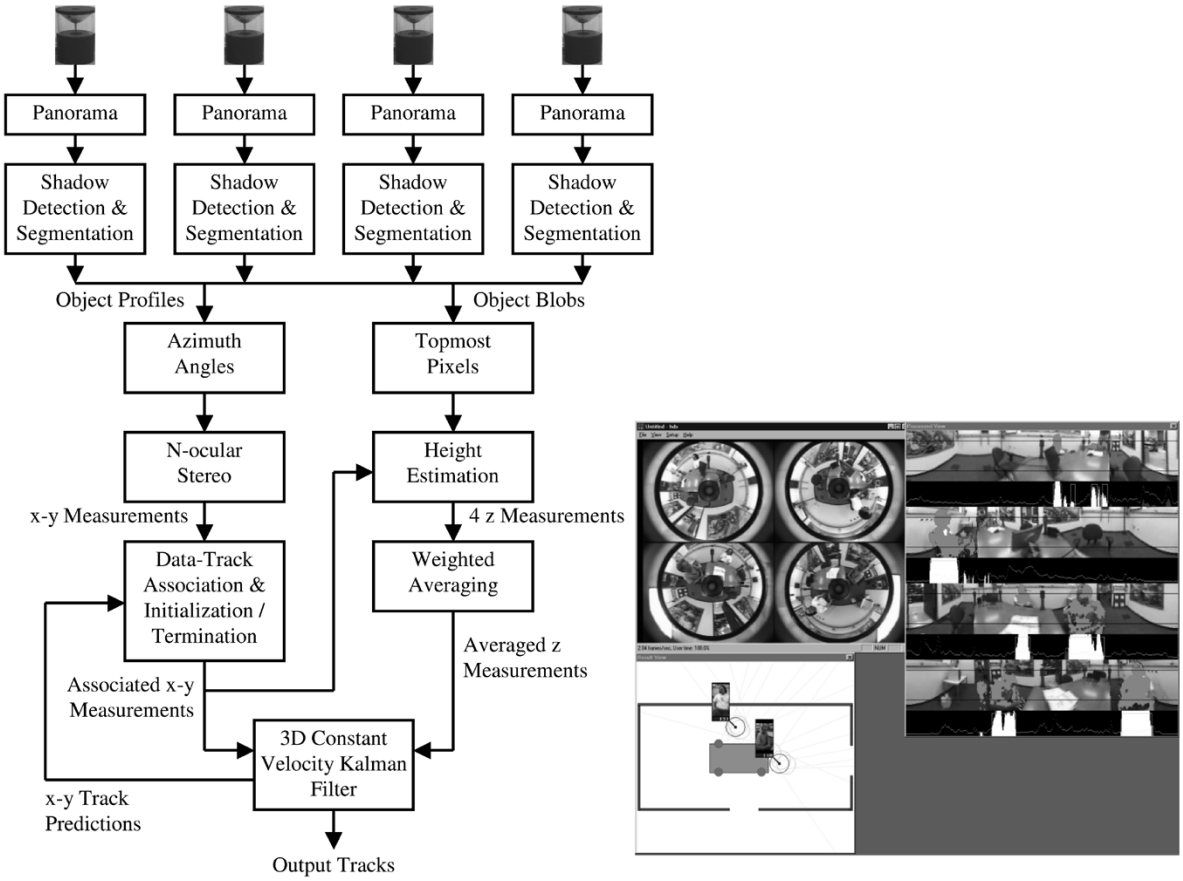


Fig. 10. Real-time 3-D O-VAT and its implementation. In the implementation, four omni videos are unwarped into panoramas where person detection and measurement are performed, and person tracking is displayed in the floor plan.

shown in the timing diagram in Fig. 8. Tracking could be running on one computer and the trajectories are communicated to other machines through network. Dynamic event capture takes one thread to compute the attentive directions to the interested low-level events. High-level event analysis takes spatial-temporal visual-audio events and derives semantic interpretations of the human activities by dynamic multistate models. Those processes achieve minimum delay and optimal efficiency by carefully synchronized multithreading.

The features of the system architecture of our current intelligent complex testbed are summarized in Table I.

IV. Tracking and Analysis of Humans in Intelligent Environments and Experimental Studies

As mentioned in Section II, video arrays deployed in an intelligent environment need to support a number of important tasks. These include: tracking of human movements, human identification, and human body analysis including gait and gesture recognition. Also, based upon the state and context of the intelligent environment, the system should be able to switch between functionalities of the video modules. In this section, we present subsystems for multiperson tracking as well as for human body analysis and give experimental results.

A. Multiperson Tracking Using Video Arrays

We have developed a real-time intelligent room system, the networked omni video array (NOVA) system as shown in Fig. 9 which utilizes the omnicam array for tracking, face capture, and face recognition [6]. It is a subsystem of Fig. 2. The 3-D tracker takes the ODVS array videos for detecting and tracking people on their planar locations as well as heights, and sends their tracks to another computer. Active camera selection (ACS) and dynamic view generation (DVG) modules in the second computer use the track information to latch upon person's face by a perspective view generated from an ODVS video in the array. Since the view is generated electronically, the face is immediately captured according to the direction of the tracker. A 64×64 face video is then extracted from the perspective view to be identified.[1] This system provides a platform for developing and evaluating robust face recognition schemes in order for the humans to behave naturally in the intelligent room.

The omnicam-based person tracker or omni-video array tracker (O-VAT) [6] is shown in Fig. 10. Silhouettes of people are detected by background subtraction with shadow removal on the panoramas unwarped by the omnicam videos. The horizontal locations of people are first measured from the azimuth angles of the silhouettes or blobs for each panorama by N-ocular stereo [28], [27] and associated to the existing

[1]Demonstration clips of person and face tracking on the ODVS array is available at http://cvrr.ucsd.edu/pm-am/demos/index.html.

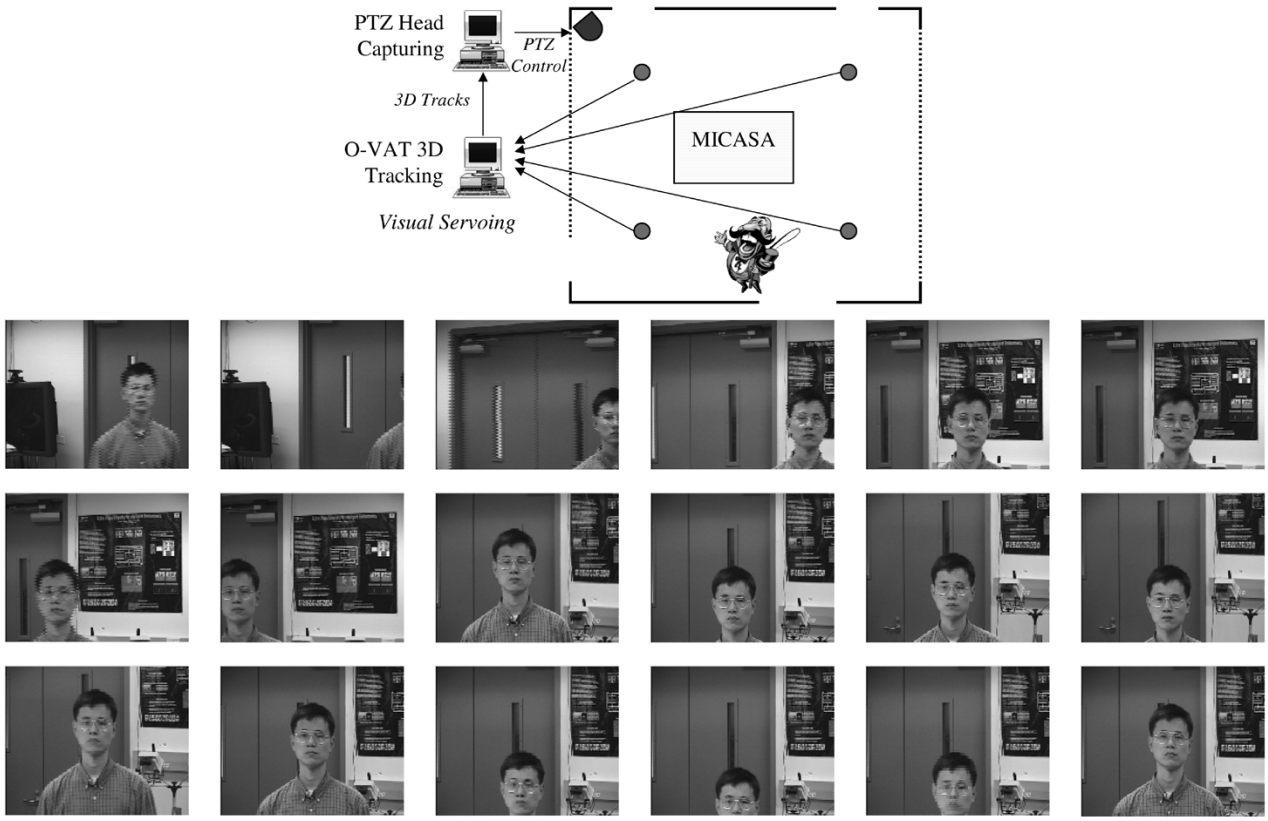


Fig. 11. Visual servoing for face capture on a mechanical PTZ camera driven by the 3-D O-VAT. In the sample sequence, please note the motion of the subject both in horizontal and vertical directions, and the dynamics of the PTZ camera trying to catch up the human motion.

set of tracks by the nearest neighborhood. If a measurement is not associated to the tracks, a new track is initialized with a time delay to avoid spurious detection. The tracks that have no measurements are also kept for some period before termination to avoid missing detection. The heights of people are measured from the topmost pixels of the silhouettes of people in the panorama by triangulation with the knowledge of the horizontal distances of people to the camera. Height measurements of one person from the cameras are averages with weights reciprocal to the horizontal distances of the person to the cameras. Then, a constant velocity Kalman filter is updated by the associated set of 3-D measurements for each person. The Kalman filter uses random maneuver with fixed maneuver covariance, but the time interval between frames is updated on-the-fly. Also, the measurement covariance is fixed and estimated empirically. They are fine-tuned for regular indoor human motions.

B. Human Face, Body, and Movement Analysis

1) Head Tracking and Face Capture: Results of the tracker are used to control the face capture module. As shown in Fig. 9, one ODVS in the array may be picked and captured in full-frame to capture the face. The advantage is that electronic PTZ is instantaneous. However, image resolution would be lower. The face capture using mechanical PTZ is shown in Fig. 11. The O-VAT uses the ODVS array on the ceiling in the MICASA testbed to track people. The location of the head is then estimated and used to drive a PTZ rectilinear camera through RS-232 commands. From the video sequence in Fig. 11, it can be seen that mechanical PTZ would have some control delay problems, and the human-motion speed needs to be limited.

There is a possible way to improve the PTZ face tracking. Given the face that O-VAT is a low-resolution tracker, the PTZ control scheme could be fine-tuned. After the PTZ camera captures the face by the direction of O-VAT, a face detector and tracker comes in to grab the face. Autonomous face tracking servo mechanism can be implemented to keep the detected face near the center of the video. If it fails to detect the face, loses track, or has a spurious detection that does not match with O-VAT, O-VAT overrides the face tracking again. If the system decides to change target, the face tracking can also be reset by the O-VAT. By this way the inaccuracy and delay due to O-VAT can be bypassed. This mechanism is to be implemented and evaluated in the future.

2) Single-Frame Face Detection and Recognition: The captured video is then processed to detect the face and extract the face video, as shown in Fig. 12. From the head tracking output, skin tone segmentation is first used to find the face candidates. Possible face images are cropped from the skin tone blobs. Those images are then classified to reject nonfaces. A simple eigenface, or principle component analysis (PCA) method is used for both face classification and single-frame face recognition [3]. We construct the PCA feature subspace with 200 face images of multiple people and face orientations taken with the perspective unwarping of omnicam videos. The output of this module is the stream of projection vectors of the face video in the PCA subspace as well as the stream of face

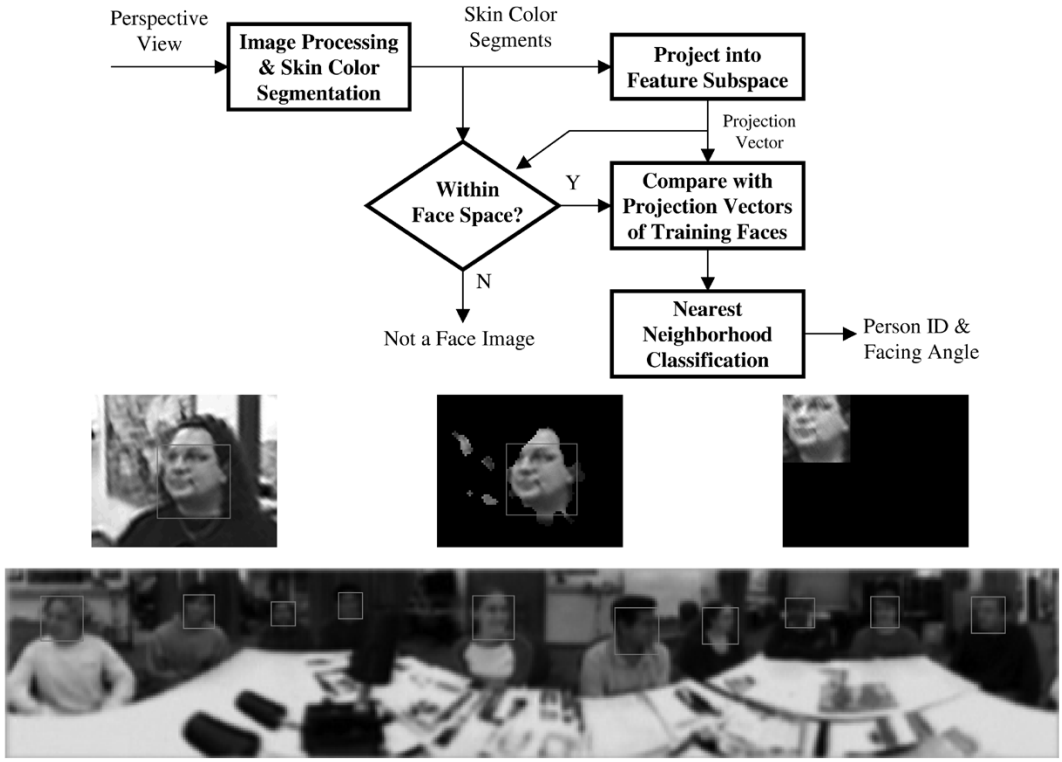


Fig. 12. Single-frame face detection and recognition on omni-vision array using view-based method. The pictures illustrate the mechanism of skin tone segmentation and face classification. Panoramic face detection is also shown.

Fig. 13. Face orientation estimation for best-view camera selection.

recognition identity. The stream of feature vectors can be further processed to estimate the face orientations and recognize a person over the frames.

As a direction of further improvement, multiple modalities of features should be used in the face detection in addition to skin tone segmentation, which works robustly only on constant illuminations like indoor environments. Possible modalities include elliptical edge links [29] and wavelets [30]. More sophisticated methods are also needed for better face/nonface classification [2], [31], [32].

3) Face-Orientation Estimation: Face-orientation estimation is needed to select a camera to capture the face with a best viewing angle. If the face capture finds a profile face, it would be necessary to capture the face by another suitable active camera. It can also be used to assess the direction of attention of people in the intelligent environment. It provides valuable information to estimate the behavior of people. We present an effective simple method built upon skin tone segmentation. Due to the hairline, the ellipse fitted to the skin pixels changes orientation as person turns from far left to far right, as shown in Fig. 13. We can regard the skin pixels as samples from a 2D Gaussian distribution and find the distribution parameters as

$$\overline{\mathbf{x}} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i,$$
$$\mathbf{C} = \frac{1}{N-1}\sum_{i=1}^{N}(\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})^T, \quad \text{where } \mathbf{x} = \begin{bmatrix} x \\ y \end{bmatrix}. \quad (1)$$

Then, the principle component, i.e., the first eigenvector corresponding to the larger eigenvalue, of the 2×2 covariance matrix $\mathbf{C}$ describes the orientation of the ellipse. A lookup table based on a set of training samples is used to relate the approximate direction the person is facing to the angle between the principle component of the ellipse and the vertical axis, as shown in Fig. 14. The table is interpolated from five facing angles of

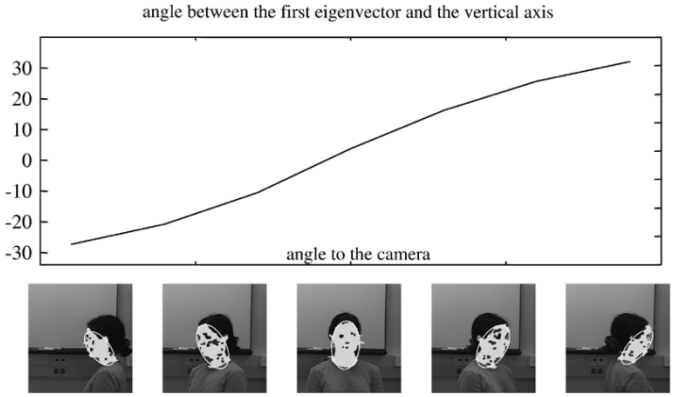


Fig. 14. Lookup table for face orientation estimation computed by averaging across the training examples.

Fig. 15. Examples of the face images in the training and testing video streams. The left six are perspective views generated from the omni videos, and the right face images are automatically extracted by the NOVA system. They show various face angles, sizes, expressions, backgrounds, and other perturbations that SFR needs to deal with.

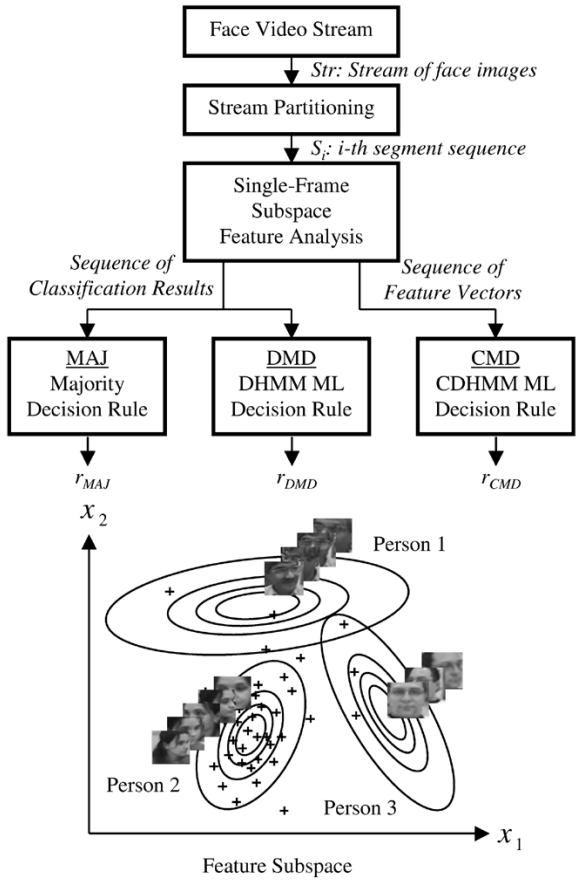


Fig. 16. Streaming face recognition scheme and the geometric interpretation in feature subspace.

the training samples. More sophisticated methods would be required for higher accuracy and robustness to cluttered backgrounds [33], [32].

4) Streaming Face Recognition: Served as a crucial event analyzer, face recognition performance can be enhanced by video-based algorithms [34]. In order to deal with uncertainties in face alignment in the captured face video, illumination changes, face orientation, gender, and racial differences, hair style and clothing, and sensor noises as shown in Fig. 15, accumulating the confidence across frames in the face video will boost the recognition accuracy. As shown in Fig. 16, the captured face video is partitioned into segments, and streams of single-frame face recognition identities as sell as PCA subspace feature vectors of the detected face are computed by subspace feature analysis module as mentioned earlier. These two streams are classified by three schemes, as shown in Fig. 16. The majority rule (MAJ) decides on the highest occurrence of the single-frame recognition identities in the segment, the discrete HMM (DHMM) maximum likelihood (ML) decision rule (DMD) decides on the sequence pattern of the single-frame recognition identities, while the CDHMM ML decision rule (CMD) decides on the stream of single-frame feature vectors.

MAJ rule is a straightforward way to decide the identity of the face video segment. It does not require training. On DMD, a DHMM is trained by the single-frame face-recognition identity streams for each person. Then, on testing phase, DMD picks the maximum value of the likelihoods of those DHMMs given the testing segment. This smoothes the jitters in the single-frame recognition sequence and would give better results than MAJ rule. However, useful information of features is already discarded by the single-frame face recognition before the DMD rule. CMD rule avoids this problem by taking the feature vector stream of subspace analysis instead of recognition identity stream. Similarly, a CDHMM must be trained for each individual, and upon testing the identity of the CDHMM that yields the maximum likelihood on the testing face video segment is decided as the final recognition output, as illustrated in Fig. 16. In the subspace interpretation of Fig. 16, each single frame of face video is represented as a point in the feature subspace, and a video segment is represented as a scatter of points. Each point has a likelihood value with respect to a specific class modeled by the Gaussian mixture density. Thus, the accumulation of the likelihoods has a maximum value if the scattering of the video segment falls mostly to the density function modeled by

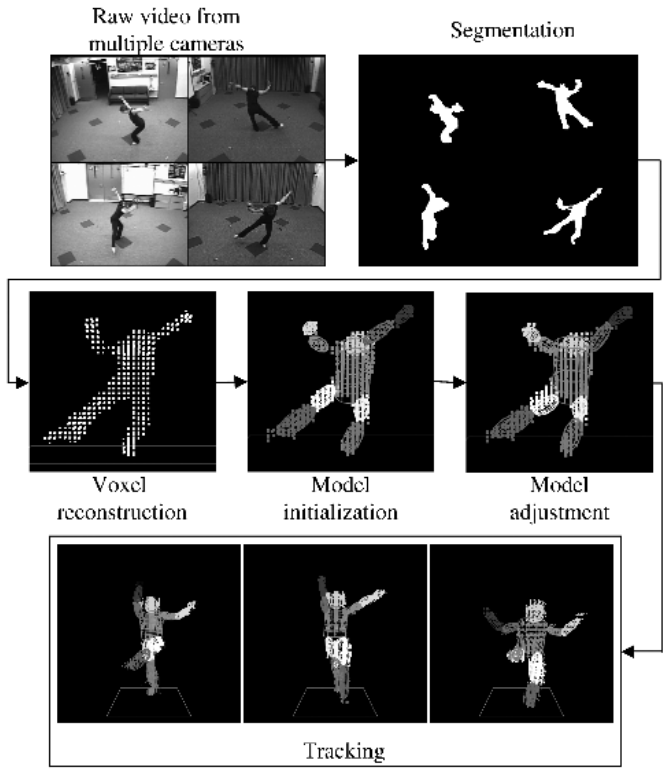


Fig. 17. Body modeling system components. Video from four cameras is segmented and 3-D voxel reconstruction is performed. Model initialization finds body parts using template fitting and growing. The positions of located body parts are then adjusted to ensure the valid model using and extended Kalman filter. A modified version of that filter is then used for tracking.

a CDHMM. This rule would give the best recognition accuracy since it is in a *delayed decision* style.

5) Body Modeling and Movement Analysis: An important feature of the MICASA testbed setup is that it allows synchronized capture of rectilinear camera array that covers an overlapped volume as shown in Fig. 6. Within the volume, voxel reconstruction of human body can be carried out in real-time. As shown in Fig. 17, the 2-D silhouettes from different viewing directions of the calibrated cameras are obtained by background subtraction with shadow detection. From the silhouettes and the calibration parameters, shape-from-silhouette is used to reconstruct the 3-D silhouette of the human. L by M by N voxels are first defined in the overlapped volume, then a voxel is marked if all the 2-D silhouettes has a pixel corresponds to it through camera calibration model. After the 3-D voxel reconstruction is available, search for the head ellipsoid is made. It is followed by fitting other ellipsoids for torso and limbs. The length and connectivity of each body parts are adjusted then by a Bayesian net. While tracking, the centroid and joints of the body parts are tracked by extended Kalman filters with respect to a body coordinate defined on the torso. Complete details of the body modeling and movement analysis system are described in a recent publication [24].[2] Some of the joint angles of the body model are plotted in Fig. 18 with two walking subjects. Differences between the walking patterns can be used as another biometrics modality for person identification, posture, and gesture analysis, and behavior recognition.

[2]The demonstration video clips are available at http://cvrr.ucsd.edu/pmam/index.html.

C. Experimental Results

In this section, we present the experimental results of some algorithms mentioned earlier.

1) 3-D Multiperson Tracking by O-VAT: Since O-VAT relates directly to the accuracy of face capturing in the NOVA system, it is necessary to evaluate its accuracy. We test it by tracking people walking on a designated path in the room around the array of four ODVS in AVIARY and in MICASA testbeds as shown in Fig. 3. People's tracks are logged for later retrieval, analysis, and plotting offline for comparison. We define the accuracy indices by the offset of the track from the ground truth and by the standard deviation of the track. In AVIARY, the four ODVSs are mounted on the four corners of a meeting table in the midst of the room and are a little higher than sitting people. The designated walking path goes around the table. Test results in AVIARY can be found in [6]. For MICASA, the room is twice the size of AVIARY, and the camera array of four ODVS is installed under the ceiling, with one ODVS near the center of each quadrant of the room (see Fig. 3). We tested O-VAT in MICASA with one to six adults as the tracking targets walking on a designated path in the room which would pass under the ODVSs. Some track plots of the targets are shown in Fig. 19, and the accuracy results are given in Table II.

From the MICASA results in Table II, the ground track offset and standard deviation increase with the number of people, especially after five people. The height estimate also degrades with the number of people. We note that for the experiments in the AVIARY testbed these indices only have a little fluctuation with one to four people [6]. The major difference between these two testbeds is that the coverage of the ODVS array in AVIARY is strictly inside-out and is about as high as people. Thus, the chance of people occluding each other is almost independent to the number of people because the ODVSs are standing upright at approximately the height of people and people walking around the array can be easily distinguished in the panoramas. To fix this occlusion problem in MICASA, the merged blobs of people in the panorama of a camera could be excluded from the measurement calculations.

O-VAT runs at approximately 20 frames per second on a platform of dual Pentium III 866 MHz, 256 MB memory, Windows 2000. Therefore it is very suitable for real-time applications. In term of system flexibility, ODVS array has good reconfigurability because with the same ODVS array, the system not only allows tracking but also allows electronic PTZ for higher level processing. Note that electronic PTZ does not require mechanical control, which leads to delay and damping problems. Also multiple electronic PTZ views at different objects can be generated from the same ODVS video at the same time. In addition, since the ODVS array can be placed in the midst of the meeting participants, it has the advantageous inside-out coverage on people's faces from a close distance by unobtrusive electronic PTZ views. Therefore ODVS network is very suitable for a meeting room setup.

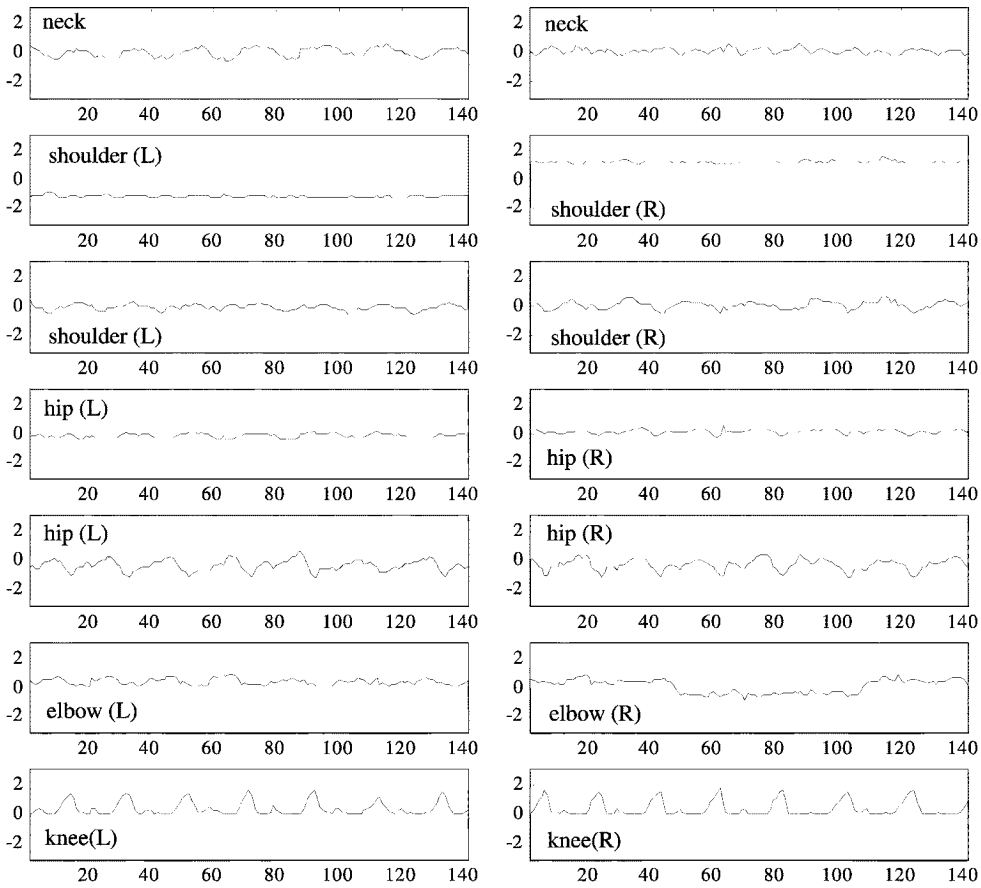


Fig. 18. Comparison of multiple joint angle patterns of the body model when tracking two walking subjects.

2) Streaming Face Recognition: For streaming face recognition (SFR) experiments, the parameters include video segment length L, the number of states N for both the DHMM and CDHMM, the number of Gaussian mixture M, and the utilized first d dimensions of the PCA feature vector of full dimension D for the CDHMM. The accuracy is evaluated as the overall correct percentage (OCP), defined as the correct recognition percentage of all frames in the single-frame case, and the percentage of correct recognized segments of all video segments in the streaming cases. Table III compares the best OCP of the recognition schemes. This outcome justifies the streaming type processing schemes because accumulating the likelihoods of the frames in a video segment would provide a better match to a class in the feature subspace than only one single frame, as illustrated in Fig. 16. It also confirms the implications of the streaming-type recognition schemes as mentioned earlier.

V. Human Activity and Interactions in an Intelligent Meeting Room (IMR)

IMR are spaces, which support efficient and effective interactions among their human occupants as in Fig. 20. They can all be occupying the same physical space or they can be distributed at multiple/remote sites. The infrastructure which can be utilized for such intelligent rooms include a suite of multimodal sensory systems, displays, pointers, recording devices, and appropriate computing and communications systems. The necessary intelligence of the system provides adaptability of the environment to the dynamic activities of the occupants in the most unobtrusive and natural manner.

The types of interactions in an intelligent environment impose requirements on the system that supports them. In an intelligent meeting room we identify three types of interactions:

1) between active participants—people present in the room;
2) between the system and the remote participants;
3) between the system and the future participants.

The first category of interactions defines the interesting events that the system should be able to recognize and capture. The active participants do not obtain any information from the system but cooperate with it, for example by speaking upon entering the room to facilitate accurate person identification. Two other types of interactions are between the system and people that are not present in the room. Those people are the real users of the system. For the benefit of the remote participant, the video from active cameras that capture important details such as a face of the presenter or a view of the whiteboard should be captured and transmitted. Information on identities of active participants, snapshots of their faces and other information can be made available. The future participant, the person reviewing the meeting that happened in the past, requires a tool that graphically summarizes past events to easily grasp the spatio-temporal relationships between events and people that participated in them. Also, an interface for interactive browsing and review of the meeting is desirable. It would provide easy access to stored information about the meeting, such as identities and

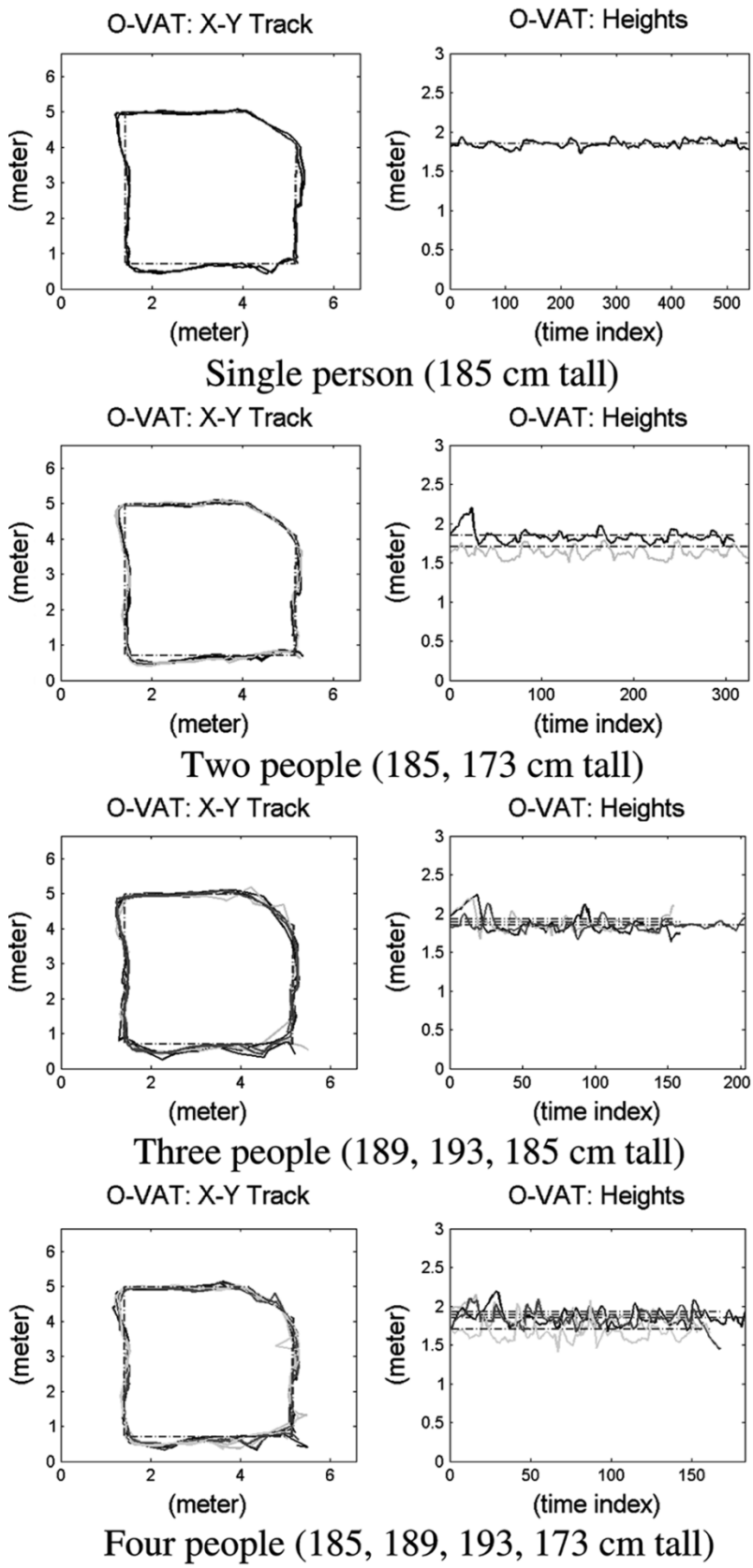


Fig. 19. Some tracking results of O-VAT in MICASA testbed with one to four adult people walking simultaneously on the designated path. The floor plan on the left shows the 2-D tracking accuracy. Dash lines are the designated walking paths on the floor (ground truth). The height tracking on different tracks of people are color-coded. The height tracking on the right are plotted against time. The actual heights of the volunteers are shown as dash lines and denoted below the plots.

snapshots of participants and video from active cameras associated with specific events.

Interactions between active participants in a meeting room define interesting activities that the system should be able to

TABLE II
TRACKING ACCURACY OF O-VAT IN THE MICASA TESTBED FOR ONE TO SIX PEOPLE. ACCURACIES ARE COMPARED IN TERMS OF TRACK MEAN OFFSET FROM THE GROUND TRUTH AND STANDARD DEVIATION OF THE TRACK, FOR BOTH X-Y GROUND TRACK AND HEIGHT ESTIMATION

Tracking Accuracy (cm)	X-Y Ground		Height Z	
	$\Delta\mu$	σ	$\Delta\mu$	σ
Single Person	11	5	1	4
Two People	10	6	2	7
Three People	12	10	3	11
Four People	13	11	5	11
Five People	19	20	5	19
Six People	29	31	9	22

TABLE III
COMPARISON OF THE BEST OCP OF THE SINGLE-FRAME FACE RECOGNITION AND THE SFR RULES. COMMON SETTINGS: $D = 135, L = 49$, NON-OVERLAPPING VIDEO SEGMENTS

Decision Rules		Best OCP	Note
Single-Frame FR		75.9 %	
SFR	MAJ	81.7 %	
	DMD	89.7 %	N=14
	CMD	99.0 %	N=1, M=1, d=8

Fig. 20. Configuring AVIARY in an Intelligent Meeting Room. People are tracked, identified, and classified as presenters or participants based upon dynamic analysis of the visual and audio signals. A summarization module maintains a record of all state changes in the system which can be accessed at a later time.

recognize and capture. We identified three: 1) a person located in front of the whiteboard; 2) a lead presenter speaking; and 3) other participants speaking. A lead presenter is the person currently in front of the whiteboard. First, activity should draw attention from one active camera that captures a view of the whiteboard. Two other activities draw attention from an active camera with the best view of the face for capturing the video of the face of the current speaker.

To recognize these activities, the system has to be aware of the identities of people, their locations, identity of the current speaker and the configuration of the room. Basic components of the system that enable described functionality are:

1) 3-D tracking of centroids using static cameras with highly overlapping fields of view;
2) person identification (face recognition, voice recognition, and integration of the two modalities);
3) event recognition for directing the attention of active cameras;
4) best-view camera selection for taking face snapshots and for focusing on the face of a current speaker;
5) active camera control;
6) graphical summarization/user interface component.

Tracking and face-recognition algorithms using visual data are already discussed in the previous section. In this section, we will also explain the role of audio data. Integration of audio and video information is performed at two levels. First, the results of face and voice recognition are integrated to achieve robust person identification. At a higher level, results of 3-D tracking, voice recognition, person identification (which is itself achieved using multimodal information) and knowledge of the structure of the environment are used to recognize interesting events. When a person enters the room, the system takes the snapshot of their face and sample of their speech to perform person identification using face and voice recognition [35], [36].

The system-block diagram is shown in Fig. 21. As mentioned before, it currently takes inputs from four static cameras with highly overlapping fields of view, four active cameras, and two microphones. All of the eight cameras are calibrated with respect to the same world coordinate system using Tsai's algorithm [27]. Two PC computers are used. One performs 3-D tracking of blob (people and objects) centroids based on input from four static cameras. Centroid, velocity, and bounding cylinder information are sent to the other PC which handles all other system functions. For new people in the environment, the camera with the best view of the face is chosen and moved to take the snapshot of the face. The person is also required to speak at that time and the system combines face and voice recognition results for robust identification. Identity of the current speaker is constantly monitored and used to recognize interesting events together with 3-D locations of people and objects and known structure of the environment. When such events are detected, the attention of active cameras is directed toward them.

The IMR project is designed for a meeting-room scenario. It not only tracks people and recognizes them, but also detects speaker activities and archive the events. The speaker activity detection is composed of a voice gate based speech detector and IBM ViaVoice speaker recognition. When a person walks in the room, the system recognizes the person by a face snapshot and speech. The identity is tagged to the track of the person. Events are defined according to the room setup. In AVIARY, an area is defined near the white board as the presenter's zone. If a person is in the area, then that person is regarded as a presenter. When the people are nearly static, the meeting starts and speech activities trigger events such as presenting, listening, and speaking (questioning/answering) and a PTZ camera zooms into the event.

A. Event Recognition for Directing the Attention of Active Cameras

This module constantly monitors for events as described in Section II. When a new track is detected in the room, it is classified as a person or object depending on the dimensions of the

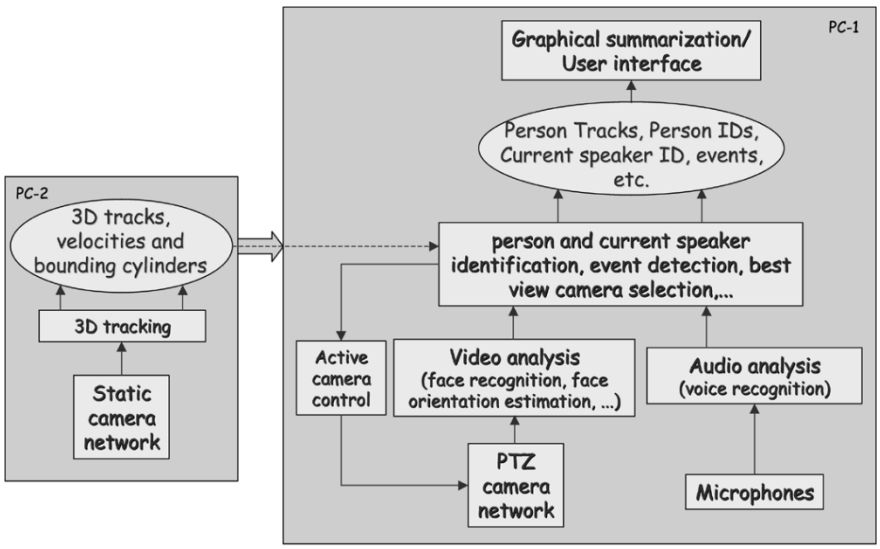


Fig. 21. Block diagram of the IMR system.

bounding cylinder. This classification is used to permanently label each track. If classified as an object, the camera closest to it takes the snapshot. If classified as a person, the camera with the best view of the face needs to be selected. The snapshot is then taken, and person identification is performed. Each person track is labeled with the person's name. Events are associated with tracks labeled as people (person located in front of a whiteboard, person in front of the whiteboard speaking, and person located elsewhere speaking) and are easily detected using track locations and identity of the current speaker.

B. Best-View Camera Selection

The best-view camera for capturing the face is the one for which the angle between the direction the person is facing and the direction connecting the person and the camera is the smallest (Fig. 22). Center of the face is taken to be 20 cm from the top of the head (which is given by the height of the bounding cylinder). There are three different situations where the best-view camera selection is performed. First is taking snapshot of the face of the person that just entered the room. Second, if the person in front of the whiteboard is speaking a camera needs to focus on their face. The third situation is when a person not in front of the whiteboard speaks. In these three situations, we use different assumptions in estimating the direction the person is facing.

When a person walks into the room, we assume that they are facing the direction in which they are walking. If a person is in front of a whiteboard (location of which is known), one camera focuses on the whiteboard (Fig. 23). If the person starts speaking, a best-view camera needs to be chosen from the remaining cameras to focus on that person's face. Since the zoomed-in whiteboard image contains person's head, we use that image to estimate the direction the person is facing by the method described in Section IV-B (Fig. 14). These estimates are not very accurate, but we have found that this method works quite reliably for purposes of best-view camera selection. In the third case, where person elsewhere in the room is speaking,

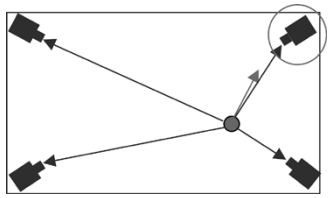


Fig. 22. Best-view camera is chosen to be the one the person is facing the most (maximum inner product between the direction the object is facing and direction toward a camera).

Fig. 23. Person standing close to the whiteboard draws attention from one active camera.

we assume they are facing the person in front of the whiteboard if one is present there. Otherwise, we assume they are facing the opposite side of the room. By these assumptions, the first image obtained with the chosen camera is processed for facing direction and the camera selection is modified if necessary.

C. Active Camera Control

Pan and tilt angles needed to bring a known location to the center of the image can be easily computed using the calibrated camera parameters. However, the zoom center usually does not coincide with the image center. Therefore, the pan and tilt angles needed to direct the camera toward the desired location have to

TABLE IV
EVENT LOG DATABASE FOR ACTIVITY ARCHIVING AND RECALL. ENTRIES ARE LOGGED WHEN THERE IS CHANGE OF THE STATES. (K = KOHSIA, I = IVANA; AND M = MOHAN)

Time Stamp	Person ID (Location)	Speech Activity	IMR state / Context (# Occupants)
0	0	0	Vacant
1	K (3.3, 0.2)	0	Occupied (1)
2	K (2.5, 0.7) I (3.3, 0.2)	0	Occupied (2)
3	K (1.5, 1.3) I (4.1, 0.5)	0	Occupied (2)
4	K (0.5, 2.1) I (5.2, 0.8)	0	Occupied (2)
5	K (0.4, 2.0) I (5.2, 0.8)	K: Presenting I: Listening	Presentation (2)
8	K (0.4, 2.1) I (5.2, 0.9) M (3.3, 0.2)	K: Presenting I: Listening M: Listening	Presentation (3)
…	…	…	…
11	K (0.4, 2.1) I (5.1, 0.8) M (5.2, 2.5)	K: Listening I: Speaking M: Listening	Discussion (3)
…	…	…	…
15	K (0.4, 2.0) I (5.3, 0.8) M (5.4, 2.9)	K: Presenting I: Listening M: Listening	Presentation (3)
20	K (0.5, 2.2) I (5.2, 0.9) M (5.5, 3.0)	0	Occupied (3)
22	K (0.6, 2.1) I (5.2, 0.8) M (5.5, 3.1)	K: Listening I: Listening M: Speaking	Discussion (3)
…	…	…	…

be corrected by the pan and tilt angles between the center of the image and the zoom center (C_x, C_y) as

$$x' = M(n)\,[x - C_x] + C_x$$
$$y' = M(n)\,[y - C_y] + C_y. \tag{2}$$

Otherwise, for large magnifications, the object of interest may completely disappear from view. A lookup table is used to select a zoom needed to properly magnify the object of interest (person's face or a whiteboard). Magnifications $M\,(n)$ are computed for a subset of possible zoom values defined by a chosen zoom step. Magnifications for other zoom values are interpolated from the computed ones. The magnifications are obtained using a slightly modified version of [37]. Two images taken with two different zoom values are compared by shrinking the one taken with the larger zoom value. The value of magnification (will be smaller than one) that achieves best match between the two images is taken to be the inverse of the magnification between the two images. The absolute magnification for a certain zoom value with respect to zero zoom is computed by multiplying the magnifications of the smaller zoom steps. The image coordinates of the zoom center is determined manually by overlaying a crosshair over the view from the camera and zooming in and out until we find a point that does not move under the crosshair during zooming.

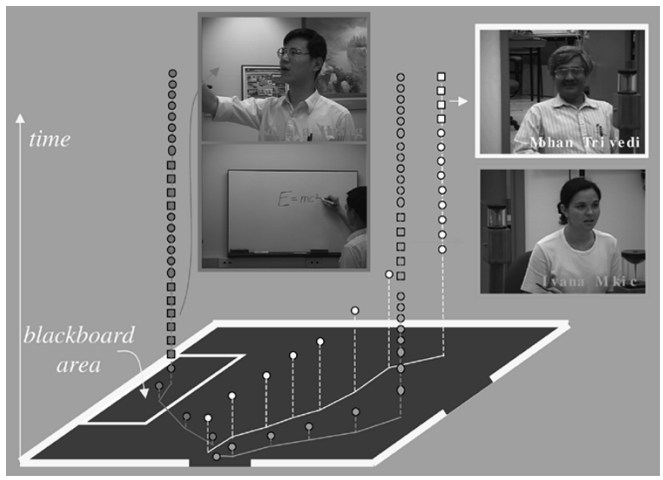


Fig. 24. Graphical summarization interface of the IMR system for retrieval. The horizontal plane is the floor plan of the meeting room, and the vertical direct represents time. People's tracks are color coded and plotted in a spatial-temporal manner. Square dots are plotted if the person is speaking, otherwise, circular dots are plotted. Interesting activities on person's location and speech activities trigger the attention from active cameras. Every object in this graphical summarization is associated with information needed to access the appropriate portion of video, face snapshots, and identity information.

D. Graphical Summarization/User Interface Module

The tracks, identities, and events are logged into a database as shown in Table IV and the audio and video are also recorded for later retrieval. A summarization interface as shown in Fig. 24 is used for the user to do the retrievals. The horizontal plane is the floor plan of the meeting room, and the vertical direct represents time. People's tracks are color coded and plotted in a spatial-temporal manner. Square dots are plotted if the person is speaking, otherwise, circular dots are plotted. Interesting activities on person's location and speech activities trigger the attention from active cameras. Every "object" in this graphical summarization is associated with information needed to access the appropriate portion of video, face snapshots and identity information. When user clicks on a circular dot, the snapshot and identity of the person is shown. If a square dot is clicked, the video clip of the speech interval is replayed. For remote viewing, the videos from active cameras that capture interesting events can be transmitted together with the other information needed to constantly update the remote summarization graph.

Experimental trials confirm satisfactory performance of the system. Person tracking module performed with maximum errors around 200 mm. These experiments included five people in the face and speaker databases, so the person identification accuracy based on both modalities is 100% in most situations. Also, recognition of the current speaker performs with nearly perfect accuracy if silence in a speech clip is less then 20% and clip is longer than 3 s. The results are very good for clips with low silence percentage even for shorter clips, but results deteriorate when silence is more than 50% of the clip. The silence percentage and speech clip length can be fine-tuned by the voice gate sensitivity. By increasing the voice gate threshold, less silence is in the speech clip, and clips less than 3 s can be discarded. However, there is an indispensable delay of 1–5 s between the beginning of speech and the recognition of the

speaker, which causes a delay in recognizing activities that are concerned with the identity of the current speaker.

The face capture and pose estimation modules work flawlessly when the moving person keeps the face in the direction of the movement. Difficulties arises when a person turns the head while walking. In this case, another camera is selected according to the estimated face orientation and the current location of the person as in Fig. 22. The camera selection for focusing on the face of the person that is talking in front of the whiteboard succeeds around 85% of the time. In the case of the person talking elsewhere in the room, our assumption that they are facing the person in front of the whiteboard or the opposite side of the room is almost always true. This is due to the room setup—there is one large desk in the middle of the room and people sit around it—therefore almost always facing the opposite side of the room, unless they are talking to the presenter. If exception happens, another camera can take over to capture the speaker on the face orientation estimations.

VI. Concluding Remarks

In this paper, we presented a framework for efficiently analyzing human activities in the environment, using networks of static and active cameras. In the framework we developed, information is extracted at multiple levels of detail depending on the importance and complexity of activities suspected to be taking place at different locations and time intervals. The environment will be constantly monitored at a low resolution, enabling the system to detect certain activities and to estimate the likelihood that other more complex activities are taking place at specific locations and times. If such an activity were suspected, to enable its accurate perception, a higher resolution image acquisition and more sophisticated analysis algorithms would be employed. The paper includes an overall system architecture to support design and development of intelligent environments. Details of panoramic (omnidirectional) video camera arrays, calibration, video stream synchronization, and real-time capture/processing are discussed. Modules for multicamera-based multiperson tracking, event detection and event-based servoing for selective attention, voxelization, and streaming face recognition are also discussed. The paper includes experimental studies to systematically evaluate performance of individual video analysis modules as well as to evaluate basic feasibility of an integrated system for dynamic context capture and event based servoing, and semantic information summarization.

The trend toward humans inhabiting intelligent or smart spaces has already begun. This will continue as high performance computing, high-speed communication links, and multi functional sensory arrays are becoming available at low cost. Integrating these modules to support natural interactions with humans in real-world situations, is still an open research problem especially from the systems engineering perspective. Satisfactory resolution of the research agenda for the development of these novel human–machine systems not only require efforts of the engineering community but also from the cognitive science, human factors and psychology communities. Such multidisciplinary efforts are already getting established, and in the not too distant future, environments such as our automobiles, highways, conference rooms, hospitals, and homes, would start displaying significant smartness in them.

Acknowledgment

The authors would like to thank the colleagues from the Computer Vision and Robotics Research Laboratory, especially N. Lassiter, K. Ng, R. Capella, and S. Cheng for their valuable contributions and collaboration. The authors would also like to thank the anonymous reviewers for their constructive comments.

References

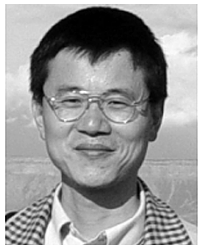
[1] R. Cipolla and A. Pentland, Eds., *Computer Vision for Human–Machine Interaction*. Cambridge, MA: Cambridge Univ. Press, 1998.
[2] E. Hjelmas and B. K. Low, "Face detection: A survey," *Comput. Vision Image Understanding*, vol. 83, pp. 236–274, 2001.
[3] A. Pentland and T. Choudhury, "Face recognition for smart environments," *IEEE Computer*, vol. 33, no. 2, pp. 50–55, Feb. 2000.
[4] R. Chellappa, C. Wilson, and S. Sirohev, "Human and machine recognition of faces: A survey," *Proc. IEEE*, vol. 83, no. 5, pp. 705–740, May 1995.
[5] B. V. K. Vijaya Kumar, M. Savvides, K. Venkataramani, and C. Xie, "Spatial frequency domain image processing for biometric recognition," in *Proc. IEEE Int. Conf. Image Process.*, vol. 1, 2002, pp. 53–56.
[6] K. S. Huang and M. M. Trivedi, "Video arrays for real-time tracking of persons, head, and face in an intelligent room," *Mach. Vision Applicat.*, vol. 14, no. 2, pp. 103–111, 2003.
[7] D. Gavrila, "The visual analysis of human movement: A survey," *Comput. Vision Image Understanding*, vol. 73, no. 1, pp. 82–98, 1999.
[8] V. Pavlovic, R. Sharma, and T. Huang, "Visual interpretation of hand gestures for human–computer interaction: A review," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 7, pp. 677–695, Jul. 1997.
[9] T. Moeslund and E. Granum, "A survey of computer vision based human motion capture," *Comput. Vision Image Understanding*, vol. 81, no. 3, pp. 231–268, 2001.
[10] L. Rabiner and B. Juang, *Fundamentals of Speech Recognition*. Englewood Cliffs, NJ: Prentice-Hall, 1993.
[11] M. S. Brandstein, "A framework for speech source localization using sensor arrays," Ph.D. dissertation, Div. Eng., Brown Univ., Providence, RI, 1995.
[12] T. Gustafsson, B. D. Rao, and M. M. Trivedi, "Source localization in reverberant environments: Modeling and statistical analysis," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 6, pp. 791–803, Nov. 2003.
[13] R. Sharma, V. Pavlovic, and T. Huang, "Toward multimodal human–computer interface," *Proc. IEEE*, vol. 86, no. 5, pp. 853–869, May 1998.
[14] V. Pavlovic, G. Berry, and T. Huang, "Integration of audio/video information for use in human–computer intelligent interaction," presented at the IEEE Int. Conf. Image Process., Santa Barbara, CA, 1997.
[15] M. Blattner and E. Glinert, "Multimodal integration," *IEEE Multimedia*, vol. 3, no. 4, pp. 14–24, Fall 1996.
[16] S. Basu, T. Choudhury, B. Clarkson, and A. Pentland, "Toward measuring human interactions in conversational settings," presented at the IEEE Int. Workshop Cues Commun., Kauai, HI, Dec. 2001.
[17] B. Brumitt, B. Meyers, J. Krumm, A. Kern, and S. Shafer, "Easyliving: Technologies for intelligent environments," in *Proc. Int. Symp. Handheld Ubiquitous Comput.*, Bristol, UK, Sep. 2000, pp. 12–29.
[18] I. Haritaoglu, D. Harwood, and L. S. Davis, "W^4: Real-time surveillance of people and their activities," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 809–830, Aug. 2000.
[19] R. Bolt, "Put that there: Voice and gesture at the graphic interface," *Proc. IEEE*, vol. 86, no. 5, pp. 853–869, May 1998.
[20] M. M. Trivedi, A. Prati, and G. Kogut, "Distributed interactive video arrays for event based analysis of incidents," in *Proc. 5th Int. IEEE Conf. Intell. Transport. Syst.*, Singapore, Sep. 2002, pp. 950–956.
[21] L. Davis, E. Borovikov, R. Cutler, D. Harwood, and T. Horprasert, "Multi-perspective analysis of human action," presented at the 3rd Int. Workshop Cooperative Distributed Vision, Kyoto, Japan, Nov. 1999.
[22] G. Cheung, T. Kanade, J. Bouguet, and M. Holler, "A real time system for robust 3-D voxel reconstruction of human motions," in *Proc. IEEE Conf. Comput. Vision Pattern Recogn.*, vol. 2, Hilton Head Island, SC, June 2000, pp. 714–720.
[23] F. V. Jensen, *Bayesian Networks and Decision Graphs*. New York: Springer-Verlag, 2001.
[24] I. Mikić, M. M. Trivedi, E. Hunter, and P. Cosman, "Human body model acquisition and tracking using voxel data," *Int. J. Comput. Vision*, pp. 199–223, 2003.

[25] N. Oliver, B. Rosario, and A. Pentland, "A Bayesian computer vision system for modeling human interactions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 831–843, Aug. 2000.
[26] S. Nayar, "Catadioptric omnidirectional camera," in *Proc. IEEE Conf. Comput. Vision Pattern Recogn.*, Jun. 1997, pp. 482–488.
[27] R. Tsai, "A versatile camera calibration technique for high-accuracy 3-D machine vision metrology using off-the-shelf TV cameras and lenses," *IEEE J. Robotics Automation*, vol. RA-3, no. 4, pp. 323–344, Aug. 1987.
[28] T. Sogo, H. Ishiguro, and M. M. Trivedi, "N-ocular stereo for real-time human tracking," in *Panoramic Vision*, R. Benosman and S. B. Kang, Eds. New York: Springer-Verlag, 2001, pp. 359–375.
[29] S. Birchfield, "Elliptical head tracking using intensity gradients and color histograms," in *Proc. IEEE Conf. Comput. Vision Pattern Recogn.*, Jun. 1998, pp. 232–237.
[30] L. Meng, T. Q. Nguyen, and D. A. Castanon, "An image-based Bayesian framework for face detection," in *Proc. IEEE Conf. Comput. Vision Pattern Recogn.*, June 2000, pp. 302–307.
[31] B. Moghaddam and A. Pentland, "Probabilistic visual learning for object representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 7, pp. 696–710, Jul. 1997.
[32] J. Ng and S. Gong, "Multi-view face detection and pose estimation using a composite support vector machine across the view sphere," in *Proc. Int. Workshop Recogn., Anal., Tracking Faces Gestures Real-Time Syst.*, Sep. 1999, pp. 14–21.
[33] K. S. Huang and M. M. Trivedi, "Robust real-time detection, tracking, and pose estimation of faces in video streams," in *Proc. Int. Conf. Pattern Recogn.*, Cambridge, UK, Aug. 2004, pp. 965–968.
[34] ——, "Streaming face recognition using multicamera video arrays," in *Proc. Int Conf. Pattern Recogn.*, vol. 4, Aug. 2002, pp. 213–216.
[35] M. M. Trivedi, I. Mikić, and S. Bhonsle, "Active camera networks and semantic event databases for intelligent environments," presented at the IEEE Workshop Human Modeling, Anal., Synthesis, Hilton Head, SC, Jun. 2000.
[36] M. M. Trivedi, K. S. Huang, and I. Mikić, "Intelligent environments and active camera networks," in *Proc. IEEE Int. Conf. Syst. Man Cybern.*, vol. 2, Oct. 2000, pp. 804–809.
[37] R. T. Collins and Y. Tsin, "Calibration of an outdoor active camera system," in *Proc. IEEE Int. Conf. CVPR*, Fort Collins, CO, June 1999, pp. 528–534.

Mohan Manubhai Trivedi (S'76–M'79–SM'86) was born in Wardha, India, on October 4, 1953. He received the B.E. (with honors) in electronics from the Birla Institute of Technology and Science, Pilani, India, in 1974 and the M.S. and Ph.D. degrees in electrical engineering from Utah State University, Logan, in 1976 and 1979, respectively.

He is a Professor in the Electrical and Computer Engineering Department, University of California, San Diego (UCSD) where he serves as the Director of the Computer Vision and Robotics Research Laboratory. His team is engaged in a broad research program in computer-vision systems, smart environments, and intelligent vehicles. He leads a number of research projects dealing with novel multisensory arrays for applications in homeland security, disaster management, and transportation infrastructure health monitoring. He is pursuing development of surveillance technologies with embedded privacy and security protections and works closely with various community, industry, and law-enforcement agencies in safety and security oriented projects. At UCSD, he also serves on the Executive Committee of the California Institute for Telecommunication and Information Technologies, Cal-$(IT)^2$, leading the team involved in the intelligent transportation and telematics projects. He also serves as a charter member of the Executive Committee of the University of California System wide Digital Media Innovation Program. He has published extensively and has edited over a dozen volumes including books, special issues, video presentations, and conference proceedings. He servers regularly as a consultant to various national and international industry and government agencies.

Prof. Trivedi is a Fellow of the International Society for Optical Engineering (SPIE). He served as an Associate Editor of the IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS for six years and was an elected member of the Administrative Committee of the IEEE SMC Society. He is a recipient of the Pioneer Award (Technical Activities) and the Meritorious Service Award of the IEEE Computer Society and the Distinguished Alumnus Award from the Utah State University.



Kohsia Samuel Huang (S'00–M'05) was born in Hsinchu, Taiwan, R.O.C., in 1966. He received the B.S. degree in electrical engineering from Chung-Yuan Christian University, Chungli, Taiwan, in 1988, the M.S. degree in control engineering from National Chiao-Tung University, Hsinchu, Taiwan, in 1991, the M.S. degree in electrical engineering from the University of Southern California, Los Angeles, in 1995, and the Ph.D. degree in electrical engineering from the University of California, San Diego (UCSD), La Jolla, in 2004.

From 1996 to 1997, he was with the Physical Optics Corporation, Torrance, CA, as a Senior Software Engineer. From 1997 to 1999, he was a Software Engineer with Chrontel, Inc., San Jose, CA. He is currently a Research Associate at UCSD. His research interests include computer vision, multimodal intelligent environments, machine learning, and signal processing.

Ivana Mikić (S'97–M'02) was born in Belgrade, Yugoslavia, in 1971. She received the B.S. degree in electrical engineering from the University of Belgrade, Belgrade, Yugoslavia, in 1994, the M.S. degree in biomedical engineering from The Ohio State University, Columbus, in 1996, and the Ph.D. in electrical engineering from the University of California at San Diego, La Jolla, in 2002.

In December 2003, she joined Vala Sciences, San Diego, CA, as a Senior Scientist. From 1994 to 1996, she was a Graduate Research Assistant at The Ohio State University. From 1996–1997, she was a Software Engineer for Motorola's Information Systems Group, Mansfield, MA. From 1997 to 2002, she was a Graduate Research Assistant at the University of California, San Diego. In the summer of 1999, she was with the Hughes Research Laboratories, Malibu, CA. In 2002 and 2003, she was a Senior Scientist at Q3-DM, Inc., San Diego, CA. Her research interests include computer vision, machine learning, and signal processing.